In-ear Audio For Physiological Monitoring

Kayla-Jade Butkow

Magdalene College
University of Cambridge

October 1, 2024

This thesis is submitted for the degree of *Doctor of Philosophy* at the Department of Computer Science and Technology

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, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

Wearable devices are revolutionising personal health and fitness monitoring by enabling real-time, continuous, and non-invasive tracking of various physiological parameters. Among wearables, *earables* (sensor-equipped earbuds) have emerged as a promising platform for physiological sensing due to their widespread use, stable position on the body, and proximity to key organs and vasculature. Research has investigated embedding specialised sensors into earables for physiological monitoring. However, earables face size and shape constraints due to their need to fit comfortably in the ears, limiting the feasibility of including such sensors. Nonetheless, commodity noise-cancelling earbuds like the Apple AirPods Pro natively contain a microphone that faces inside the ear canal. In this thesis, we explore the potential of using this in-ear microphone for monitoring physiological parameters. Specifically, we focus on three key vital signs: heart rate, respiratory rate, and stroke volume. For each of these, we develop novel sensing pipelines that leverage the unique properties of in-ear audio, including its ability to capture both heart sounds and footstep sounds.

First, we study heart rate monitoring under conditions of daily life, encompassing when the user is both sedentary and active. We propose a pipeline to perform supervised denoising of in-ear audio to emphasise heart sounds within motion-corrupted signals from which heart rate can be determined. We thus show the feasibility of accurately estimating heart rate under various conditions, demonstrating the potential for a new sensing modality for heart rate on earables.

Our second contribution explores the possibility of measuring respiratory rate, another key physiological parameter, under daily life conditions using only in-ear audio. We use physiological couplings between cardiovascular activity, gait, and respiration to indirectly estimate respiratory rate from heart sounds and footstep sounds captured by the in-ear microphone. In doing so, we effectively overcome the shortcomings of requiring reliable respiratory sounds for respiratory rate measurements. This contribution proves the possibility

of achieving robust respiratory rate estimation under various conditions using earables.

Finally, we investigate the potential for monitoring stroke volume, a clinical vital sign typically measured only in medical settings, using in-ear audio. We employ self-supervised learning and transfer learning to demonstrate the feasibility of estimating average stroke volume using only sensors available in commodity earbuds.

This thesis contributes to the growing field of earable computing and physiological sensing, advancing the understanding and application of in-ear audio for physiological monitoring. Our research demonstrates that in-ear microphones can effectively monitor key physiological parameters in real-life conditions, paving the way for widespread, continuous health monitoring. Our findings have significant implications for personal health monitoring, fitness tracking, and potential clinical applications, demonstrating the potential of earables in transforming how we monitor and understand human physiology in daily life.

Acknowledgements

I dedicate this thesis to my family, to all the generations before me who felt so strongly about the value of education. You instilled the values of academic excellence in me from such a young age that it was never a surprise that I ended up doing a PhD even though I insisted I never would. Mom and Dad, you made me the person I am today. You nurtured a love for learning and for problem-solving, an attitude of resilience and perseverance. Thank you for always supporting and motivating me. Paigie, Brenty, Ari, Cody, you are my best friends and I couldn't have done this without your love, jokes, memes, Loki photos and never-ending encouragement. I love you all. A special mention has to go to my incredible grandparents Ella and Roy who have always been my biggest cheerleaders, and to Molly and Hymie who didn't have the chance to see me grow up but are always there in my heart.

First and foremost, I want to thank my supervisor, Cecilia Mascolo. Thank you for believing in me, for nurturing me, for pushing me to be the best researcher I could be, and for making me think and critically analyse. Thank you for giving me room to grow as a researcher, for helping me to identify problems that matter and for your responsiveness and support which truly make a PhD easier. Thank you for creating a research group that makes research interesting and inspiring.

The Mobile Systems group is truly a special place, filled with learning, growing, collaboration and fun. To the Earables team, Yang, Jake, Qiang and Mathias, working with you has been the highlight of my PhD. I have loved problem-solving with you, our sometimes torturous data collections, and learning and growing from each one of you. Thank you for being the best collaborators and for all the guidance. To Andrea, my first collaborator, internship supervisor and mentor, you made my first year of PhD when I was stuck in South Africa infinitely easier through your guidance and friendship. I can't thank you enough for that. To Ian, Young, Tong, Sotiris, Yvonne, Evelyn, Jing, thank you for the informal guidance on my work and for making the office so much fun. Our lunchtime Exploding Kittens has

got me through some difficult PhD days.

I am forever grateful to Nokia Bell Labs, the Cambridge Trust, and the School of Technology whose funding made this PhD possible for me. I am also grateful to Philip Mair who moved the earth to find me PhD funding even when I thought it would be impossible. Thank you to Magdalene College and especially to the Magdalene Boat Club for helping me discover a love for rowing that I didn't ever expect to find. I will never forget Bumps in the snow or the peace of being out on the river at sunset.

Thank you to all my special friends at home who have stuck with me through this process. Sarah, Shani, Caylee, Jess, Shiann, Rafi, distance and time never change our friendship, and I appreciate each one of you so much.

Being away from home is always tough, but I've been so lucky to make friends that have turned into family. It is the people that make a home, and you have all made Cambridge into a home for me. Rachel and Jan (the dog playing society), who would have thought that living together for 6 months would have turned into a lifelong friendship. Thank you for always making me smile. Swetha, Haseeb, Erika, Lawrence and Matt, Cambridge would have been a much more miserable place without you. Thank you for the holidays, the walks, the laughter, the advice, the encouragement and for supporting me always. You will always have such a special place in my heart.

Finally, to Hayden. Thank you for supporting me through the tears, the paper rejections, the deadlines, and the stress. Thank you for making me laugh and smile even on the worst day. Thank you for being my best friend and for making me so happy. I can't wait to see what the next chapter holds for us. I love you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In recent years, the proliferation of wearable devices has started to revolutionise the landscape of personal health and fitness. With the rapid growth of these devices, we now have access to a wealth of data from the human body, collected from various locations (*e.g.,* wrist, finger, ear) and using various sensors (*e.g.,* inertial measurement unit, photoplethysmography, microphone).

Initially, this data was mainly used for basic step counting and activity tracking. However, the field has progressed to monitoring parameters related to health and physiology. Recently, wearable devices have begun to be used for heart rate monitoring, sleep tracking, and even measuring blood oxygen saturation.

Wearables are advantageous for physiological monitoring as they enable real-time, continuous and non-invasive tracking. This empowers users to better understand their health and wellbeing, customise training programs to reach health-related goals, and can even be used for early diagnosis and to track disease progression. However, the usefulness of wearables is often hampered by their form factor. For example, respiration and heart rate are typically measured using a chest strap which is inconvenient for long-term use and impedes daily activities [1]. Other modalities, such as smartwatches, are commonly used, but their measurements are prone to noise due to movement [2].

The form factor of wearable devices has evolved significantly over time. While they initially

started as smartwatches, there has been a rise in ring-based and head-based wearables over time. Head-worn devices include glasses, VR headsets, brain-computer interfaces, and more recently, earbuds. These devices offer interesting and unique opportunities for sensing as they lie close to key organs and vasculature, such as the brain, nose, mouth, and the arteries supplying the brain with blood. Among these head-worn devices, earbuds are particularly interesting. Sensor-equipped earbuds, or *earables*, currently make up the largest share of the wearables market [3], with this trend expected to continue growing year on year. Earables are unique among head-based wearables as they are already pervasive due to their use for entertainment and remote working. Besides their widespread nature, earables also have several advantages over other wearables used for health and fitness monitoring:

1. The head is a stable location on the body, making signals collected from the ear more robust than those collected from smartwatches, which suffer from random motions due to wrist and hand movements.

2. The head lies close to key organs and vasculature, enabling sensing opportunities not accessible to wrist or finger-worn devices.

3. Unlike other head-worn devices, earables are small and light enabling them to be worn for long periods, both while stationary and moving, without obstructing daily activities.

Sensors such as electroencephalogram (EEG) [1] and photoplethysmography (PPG) [4] have been embedded into earables to achieve physiological sensing applications, such as heart rate monitoring and sleep tracking [5]. However, equipping earables with additional sensors adds computational and design complexity, extra power consumption and higher cost, which hampers its potential for widespread adoption and use. Commercially available earables, such as the Apple AirPod Pro [6] and Sony WF-1000XM5 [7] are equipped with an inertial measurement unit (IMU), and internally and externally facing microphones. These sensors help to achieve fundamental functionality of the earbuds, such as automatic pause when the earbud is taken out of the ear (IMU), speech and calls (external microphone), and active noise cancellation (internal microphone). This internal microphone[1], which faces inside the ear canal, opens up new opportunities for passively sensing internal biosignals and sounds.

In this thesis, we explore the untapped potential of earables and in-ear audio sensing to

[1]This thesis uses internal microphone, inward-facing microphone and in-ear microphone interchangeably depending on context

advance the field of non-invasive physiological monitoring.

1.2 Limitations and Challenges in Earable Research for Physiological Monitoring

Earables are still in their infancy within the larger wearables ecosystem, specifically for physiological monitoring. Regardless, they have huge potential for widespread use in this area. However, several challenges and limitations in existing research need to be addressed to achieve this goal.

Sensing Modalities: Since earables are a new technology, limited research has been conducted on measuring physiological parameters with them. It remains unclear which elements of human physiology can effectively be measured with which sensing modality. It is also unclear whether the native sensors onboard commercial earables are sufficient to sense physiology or if additional specialised sensors, such as PPG, are required. Additionally, due to the complex nature of human physiology, physiological and bio-signals are inherently noisy and often have very small amplitudes. This results in low signal-to-noise ratios (SNR) and low-quality data. A key challenge in earable sensing for physiology is how to develop processing pipelines to generate meaningful measurements from this low-quality data.

Data Processing and System Performance: Earables have a small form factor due to their placement on the body, resulting in small batteries with low power capacities. Even without additional sensing requirements, earbuds typically operate for only 4 to 5 hours continuously without needing to be charged. However, to monitor physiology, a continuous stream of sensor data is needed, which requires further processing and places additional strain on the already limited battery. Additionally, the more physiological parameters one aims to measure, the higher the power requirements will be. Thus, a key challenge lies in developing data processing pipelines that can process this data with low latency to provide real-time measurements to the user, while minimising power consumption and battery usage.

Dataset Size and Generalisability: Due to the relatively early stage of the earables field, there is a lack of hardware available to collect data for developing and evaluating algorithms. Consequently, a key limitation in earables research is the development of algorithms using limited data. This presents a significant challenge in creating algorithms that can generalise

to new users within the constraints of a limited dataset.

In this thesis, we focus on sensing using the in-ear microphone. We present three physiological parameters, or vital signs, that can be measured using earables: heart rate, respiratory rate, and stroke volume. The rest of this section will relate the broad challenges facing earables for in-ear microphone sensing and the sensing of these three vital signs.

In-ear microphone sensing

Although commodity active-noise cancelling earables contain in-ear microphones, current earable platforms do not provide access to an application programming interface (API) to access this data stream. This makes accessing data from the in-ear microphone challenging. This also means that custom hardware is needed to collect in-ear audio, thus hampering the scalability of data collection exercises. Additionally, due to the high sampling rate required to capture meaningful audio signals, performing power and compute-efficient audio operations is more difficult than with other sensors. It is therefore critical to develop efficient processing techniques to minimise the power consumption of the processing pipelines. Another key challenge with in-ear audio signals is their high variability between participants. Therefore, it is essential to develop systems that can generalise well across different in-ear audio signal properties.

Heart rate monitoring

Research is underway to determine the extent to which earables can be used to measure heart rate. Research has seen PPG [4, 8] and electrocardiogram (ECG) [9] embedded into earables to measure heart rate. However, these sensors are not available on commodity earbuds. Early works explored the feasibility of using in-ear microphones to capture heart signals while stationary [10], however even while stationary, the SNR of these signals was weak and highly prone to noise. At the time of writing this thesis, no studies had explored heart rate monitoring under motion conditions using in-ear audio. As we will discuss in Chapter 3, motion conditions bring new challenges since interference from motion is difficult to distinguish from the biosignal due to large frequency overlaps, worsening the already low SNR of the signal. We also need to create models with good performance using limited-size datasets. To achieve this, we require intelligent techniques that combine deep learning and signal processing to enhance model generalisability and achieve robust results. Another key challenge is ensuring that heart rate can be measured without disrupting the primary

functionality of earables, such as music listening. Finally, from a systems perspective, it is key to develop algorithms that can determine heart rate in pseudo-real-time to provide continuous heart rate measurements to the user.

Respiratory rate monitoring

Breathing sounds are perhaps the most intuitive physiological sounds to measure using microphones since they are sometimes of audible volume. This has been leveraged to measure breathing sounds using microphones placed under the nose [11] and with smartphone microphones [12]. However, a key problem with microphone sensing is background noise. To distinguish the signal from this noise, high-intensity breathing sounds are required, which are uncommon during normal, resting breathing. Early works have also investigated capturing breathing sounds using the in-ear microphone [10], showing that ear breathing signals are weak with very low SNR. Due to the weakness of these breathing sounds, even the smallest head movements while stationary cause inaccurate estimations and noise in the signal. Due to the significant challenges imposed by motion artefacts on weak signals (as discussed in Chapter 4), at the time of writing this thesis, no studies had examined respiratory rate estimation under real-life conditions (sedentary and active) using in-ear microphones. Thus, developing signal-processing pipelines capable of accurately operating under multiple conditions remains a challenge. Additionally, for respiratory rate estimation to be performed in real-life conditions, processing must occur in pseudo-real-time to provide the user with meaningful data. This presents another challenge in developing powerful, yet efficient algorithms for this complex problem.

Stroke volume monitoring

While early works have proven the feasibility of heart and respiratory rate monitoring using earables while stationary, no studies have extended research into more clinical-based physiological parameters. This is a key limitation in current earables research that needs to be addressed. A significant challenge with this line of research is obtaining accurate ground truth due to the need for medical expertise and hospital-grade devices. One critical clinical physiological parameter is stroke volume, the volume of blood pumped by the left ventricle in one contraction. Measuring stroke volume is essential for evaluating cardiac function and assessing overall cardiovascular health and fitness. It has been shown that stroke volume can be calculated using seismocardiogram signals (heartbeat-induced chest vibration signals) [13] and PPG on a chest-worn sensor [14], and PPG on the finger [15]. However, as

we will show in Chapter 5, stroke volume has yet to be measured using commodity wearable devices. Therefore, a key challenge lies in proving the feasibility of not only measuring stroke volume from the ears but also doing so using an untested sensing modality. Another significant challenge in this line of work is the limited dataset size. Because measuring stroke volume requires clinical staff, the number of participants is further restricted. This necessitates the use of learning-based techniques to extract key features from the data, enabling generalisation even with a limited dataset.

1.3 Thesis and Substantiation

We have seen that research on earables for physiological sensing is still in its early stages, particularly when using the sensors natively available on commercial earbuds. Of these sensors, the in-ear microphone shows promise for monitoring human physiology. However, its use is coupled with numerous challenges and limitations that must be addressed using intelligent processing techniques. To address these, our thesis aims to: *enable in-ear audio for physiological monitoring by exploring biosignals captured by the in-ear microphone and developing intelligent signal processing and learning-based systems to monitor key aspects of human physiology.* We substantiate this statement by developing systems to monitor three key physiological parameters.

Ultimately, this thesis addresses the following research questions:

- Research Question 1: How can we develop pipelines to overcome low signal-to-noise ratio and high motion interference in in-ear audio signals to determine heart rate?
- Research Question 2: How can we leverage in-ear audio and its unique properties to estimate respiratory rate even under inaudible breathing sounds?
- Research Question 3: How can we develop high-performing models to measure stroke volume from limited in-ear audio datasets?

1.4 Contributions and Chapter Outline

This thesis will start with an overview of the background and existing works in earable sensing and in-ear audio sensing in Chapter 2, before presenting the three main contributions which address the research questions posed in the previous section:

Contribution 1: Motion-resilient heart rate monitoring

In Chapter 3, we investigate heart rate monitoring using in-ear audio. Specifically, we assess the feasibility of monitoring heart rate under real-life conditions, including both sedentary and moving scenarios. We leverage audio collected from inside the ear canal, which contains heart sounds due to the occlusion effect (the amplification of low-frequency bone-conducted sounds inside the occluded ear canal). While this audio includes heart sounds, these are often corrupted by motion artefacts when the user is moving. To address this issue, we propose hEARt, a deep learning-based pipeline for motion-resilient heart rate estimation. In this pipeline, we use a U-Net model to denoise and enhance in-ear audio using ECG signals followed by a signal processing pipeline to estimate heart rate. We further pre-train the model with a heart sounds dataset to initialise the model weights. We implemented the system using a custom earable device with an in-ear microphone in each ear and collected data from 15 participants under various sedentary and motion conditions. Our results show mean absolute errors (MAE) of 1.88 beats per minute (BPM), 6.83 BPM, and 13.19 BPM for sedentary, walking, and running, respectively, with errors of less than 10% across all activities. Furthermore, our in-ear audio-based heart rate estimation outperforms PPG-based measurements on earbuds. We also demonstrate that heart rate can be measured while using the essential functionalities of earbuds, such as speaking and listening to music. We demonstrate, for the first time, the potential of using in-ear audio for heart rate monitoring in real-life conditions.

Contribution 2: Robust respiratory rate monitoring

Previously, we identified that in-ear audio signals exhibit clear heart sounds when stationary and clear footstep sounds when moving. We build upon these findings in Chapter 4 and investigate whether it is possible to monitor respiratory rate while sedentary and active. We found that breathing sounds cannot be reliably detected using the in-ear microphone during resting conditions which are dominated by light, low amplitude breath sounds, or under active conditions where they are overwhelmed by footstep sounds. To achieve robust monitoring, we leverage the unique ability of the in-ear microphone to capture both heart sounds and footstep sounds. Specifically, we propose RespEar, a system for indirectly estimating respiratory rate using Respiratory Sinus Arrhythmia (RSA) and Locomotor Respiratory Coupling (LRC), which are physiological couplings between cardiovascular activity, gait and respiration. The system consists of two parallel signal processing pipelines

based on user activity: one for sedentary conditions, extracting respiratory rate from heart rate variability signals, and another for active conditions, extracting respiration-related features from in-ear audio combined with stride frequency to estimate respiratory rate. Using our earable prototype, we collected data from 18 participants across 8 different scenarios. We achieved a MAE of 1.48 breaths per minute (BPM)[2] and a mean absolute percent error (MAPE) of 9.12% in sedentary conditions, and an MAE of 2.28 BPM and a MAPE of 11.04% in active conditions. We demonstrate the feasibility of using a single commodity device, earbuds, to measure respiratory rate across various daily life activities.

Contribution 3: Stroke volume monitoring

Building on the clear heart sounds captured while sedentary in the previous two chapters, we study the feasibility of monitoring stroke volume using in-ear audio. Stroke volume is a key marker of cardiovascular health and physical fitness, but it is typically measured only within clinical settings. We demonstrate the possibility of measuring this clinical vital sign using only sensors on commodity earbuds. Specifically, we use generative self-supervised learning with a masked autoencoder and transfer learning to predict average stroke volume before and after isometric exercise. Since the signal properties of in-ear audio are vastly different from those of general audio, we compiled a dataset of 1160 minutes of unlabelled in-ear audio from various participants and activities and used this dataset to train the encoder-decoder network, adapting it to in-ear audio. We then remove the decoder network, and further fine-tune the encoder on a specialised dataset specifically collected for stroke volume estimation. With data from 23 participants, we achieve an overall MAE of 5.24 mL (6.83%), with errors of 6.81% and 6.85% before and after exercise, respectively. We thus demonstrate the feasibility of measuring stroke volume using commodity wearables, paving the way for widespread monitoring of cardiovascular health and fitness.

1.5 List of Publications

This section contains the works that have been submitted to and published in peer-reviewed journals, workshops and conferences, both directly related to this dissertation and otherwise.

[2]Both heart rate and respiratory rate errors are expressed in units of BPM, with BPM meaning beats per minute for heart rate, and breaths per minute for respiratory rate. In this thesis, we will use BPM for both with the meaning being implied by context.

Works related to this dissertation

[16] **An evaluation of heart rate monitoring with in-ear microphones under motion**
Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, Yang Liu, Cecilia Mascolo.
Pervasive and Mobile Computing

[17] **hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones**
Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, Cecilia Mascolo.
2023 IEEE International Conference on Pervasive Computing and Communications (PerCom), Atlanta, GA, USA, 2023

[18] **RespEar: Earable-Based Robust Respiratory Rate Monitoring**
Yang Liu[3], Kayla-Jade Butkow[3], Jake Stuchbury-Wass, Adam Pullin, Dong Ma, Cecilia Mascolo
2025 IEEE International Conference on Pervasive Computing and Communication (PerCom), Washington DC, USA, 2025

[19] **Measuring Cardiac Stroke Volume Through In-ear Audio Sensing**
Kayla-Jade Butkow, Navazh Jalaludeen, Yang Liu, Jake Stuchbury-Wass, Mathias Ciliberto, Dong Ma, Joseph Cheriyan, Cecilia Mascolo
Submitted

Other Works

SmarTeeth: Augmenting Manual Toothbrushing with In-ear Microphones
Qiang Yang, Yang Liu, Jake Stuchbury-Wass, Kayla-Jade Butkow, Dong Ma, Cecilia Mascolo.
CHI '25: CHI Conference on Human Factors in Computing Systems

WalkEar: An Earable-based Application-agnostic Gait Monitoring System
Jake Stuchbury-Wass, Yang Liu, Kayla-Jade Butkow, Josh Carter, Qiang Yang, Ezio Preatoni, Dong Ma, Cecilia Mascolo.
2025 IEEE International Conference on Pervasive Computing and Communication (PerCom), Washington DC, USA, 2025

[3]Both authors contributed equally to the research work reported in this paper.

[20] **EarTune: Exploring the Physiology of Music Listening**
Kayla-Jade Butkow, Andrea Ferlini, Fahim Kawsar, Cecilia Mascolo, Alessandro Montanari.
In Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '24).

[21] **BrushBuds: Toothbrushing Tracking Using Earphone IMUs**
Qiang Yang, Yang Liu, Jake Stuchbury-Wass, Kayla-Jade Butkow, Dong Ma, Cecilia Mascolo.
In Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '24).

EarMeter: Continuous Respiration Volume Monitoring with Earable Audio
Yang Liu, Qiang Yang, Kayla-Jade Butkow, Jake Stuchbury-Wass, Dong Ma, Cecilia Mascolo
Under Review: MobiCom 2025

Deep-Learning Based Segmentation of In-Ear Cardiac Sounds
Jordan Waters, Jake Stuchbury-Wass, Yang Liu, Kayla-Jade Butkow, Cecilia Mascolo.
Under Review: Artery Research special collection on Wearable devices for BP and hemodynamics

[22] **IMChew: Chewing Analysis using Earphone Interial Measurement Units**
Tamisa Ketmalasiri, Yu Yvonne Wu, Kayla-Jade Butkow, Cecilia Mascolo, Yang Liu.
Proceedings of the Workshop on Body-Centric Computing Systems, Minato-ku, Tokyo, Japan, 2024

[23] **Heart Rate Extraction from Abdominal Audio Signals**
Jake Stuchbury-Wass, Erika Bondareva, Kayla-Jade Butkow, Sonja Šćepanović, Zoran Radivojevic and Cecilia Mascolo.
ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.

Patents

Yang Liu, Kayla-Jade Butkow, Dong Ma, Cecilia Mascolo. RESPIRATORY RATE SENSING. GB2314981.8, 2023.

Chapter 2

Related Work

In the previous chapter, we explored the potential of in-ear audio for physiological monitoring and discussed the challenges and limitations of physiological monitoring using earables. In this chapter, we proceed by reviewing prior techniques and applications of earables and wearables to physiological monitoring, as well as common techniques for audio signal processing. Specifically, we briefly discuss general applications of earables to health and wellbeing in Section 2.1, followed by an introduction to passive acoustic sensing on earables in Section 2.2. We then focus on the background and prior work of the three application scenarios of interest in this thesis: heart rate (Section 2.3), respiratory rate (Section 2.4) and stroke volume (Section 2.5). Finally, we discuss common processing techniques for audio signals, including signal processing and deep learning, in Section 2.6.

2.1 Health and Wellbeing Monitoring Using Earables

Recently, the earable research field has seen significant growth, with a substantial amount of work being published. There has been a particular focus on using earables for monitoring health and wellbeing, especially in the areas of cardiac and respiratory sensing. This will be elaborated on in Sections 2.3.2 and 2.4.2.

In addition to cardiac and respiratory monitoring, earables have been used to measure other physiological parameters, including blood oxygen saturation [24, 25], body temperature [26], blood pressure [27], and VO_2max [28]. Additionally, earables have been used for the detection of bruxism [29], swallowing detection [30, 31], chewing [32] and food intake

analysis [33]. Research has also explored using earables for cough detection [34], bowel sound detection and analysis [35], and the detection and differentiation of ear diseases [36].

Therefore, it is evident that there is growing interest in using earables to monitor health and physiology, making them a promising tool for cardio-respiratory monitoring.

2.2 Passive Acoustic Sensing on Earables

The works discussed in Section 2.1 leverage various sensors, such as IMUs, external microphones or PPG on earables to achieve diverse sensing capabilities. However, of interest in this thesis is the inward-facing microphone. In this section, we provide a primer on the use of the inward-facing microphone for sensing and discuss related works that intelligently leverage this microphone for various applications.

2.2.1 Enabling passive acoustic sensing

To achieve active noise cancellation (ANC), ANC earbuds are equipped with both an inward-facing microphone, facing inside the ear canal, and an external microphone, facing the environment. The inward-facing microphone functions as a feedback microphone, working together with the external microphone to remove external noise, thereby achieving noise cancellation [5]. Aside from their primary function in ANC, inward-facing microphones also enable passive sensing capabilities. This opens up new opportunities for sensing sounds inside the ear canal, thus enabling the capture of body sounds that were previously inaccessible.

However, these body sounds are weak and difficult to capture. Therefore, to enable meaningful signal capture using the inward-facing microphone, we leverage a key physiological phenomenon known as the *the occlusion effect*

When the ear canal is occluded, or blocked, the impedance within the ear canal increases. This causes the ear canal to act as a low-pass filter, amplifying low-frequency sounds while attenuating high-frequency sounds [37]. The occlusion effect manifests specifically with bone-conducted sounds, which are sounds conducted through the bones to the ear [38]. These bone-conducted sounds travel through the bones to the inner ear and cause vibrations in the walls of the ear [37]. When the ear canal is open, these sound waves escape through the ear canal. However, when the canal is blocked, the sound waves reflect off the occlusion

device and onto the eardrum where they are amplified [39]. Research has shown that the occlusion effect can cause amplifications of up to 45 dB at 150 Hz, with amplification decreasing to 10 dB at 1000 Hz [40].

These bone-conducted sounds are typically related to physiological noises, such as heart sounds, breathing, footsteps, voice, and chewing noises [10, 40]. Of particular interest in this thesis is the bone-conducted sounds associated with heartbeats, breathing and steps. Earables can take advantage of the occlusion effect by blocking the ear canal, allowing for the capture of these sounds. Therefore, using earables that occlude the ear canal, we can leverage this effect to record bone-conducted sounds through the inward-facing microphone, as shown in Figure 2.2.1.

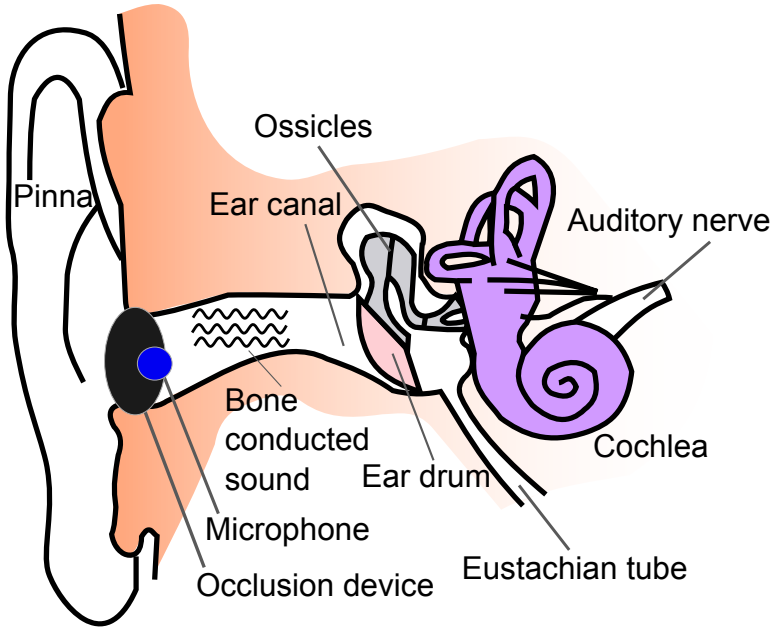


Figure 2.2.1: The anatomy of the ear and the occlusion effect. Adapted from [41].

2.2.2 Applications of passive acoustic sensing

The occlusion effect in earables has been used for various sensing applications. Ma et al. [42] employed the occlusion effect to detect various human activities, such as walking, running, drinking and remaining stationary, and various hand-to-face gestures.

This study also demonstrated the potential for detecting steps during walking and running and showed that accurate step counting can be done using earables, thus proving that in-ear

audio signals can capture clear step sounds during motion. Additionally, passive sensing has also been used to distinguish between different activities of the teeth [43], recognise different hand-to-face gestures [44], monitor eating and drinking activities [33], and provide biometric authentication [44–46]. Other applications include sleep sound classification [47], sleep staging [48], and spirometry [49].

In the coming sections, we will present the background and related work of the three specific physiological phenomena we focus on in this thesis.

2.3 Heart Rate Monitoring

Heart rate is an excellent indicator of fitness level, and is strongly associated with cardiovascular disease and mortality risk. Heart rate monitoring can help design workout routines to maximise training effect, and, more importantly, serves as an early biomarker for heart disease since cardiovascular fitness is a key predictor of cardiovascular disease. Additionally, heart rate variability (HRV), which measures the variation in time between successive heartbeats, is a predictor of physical and mental health. HRV, a proxy for autonomic nervous system behaviour, is predictive of aerobic fitness when measured during both maximal and sub-maximal exercise [50]. Thus measuring heart rate during physical activity is critical for monitoring human health and wellbeing. In this section, we first discuss general wearables for heart rate monitoring to provide context on the shortcomings of existing wearables. Then we discuss earable-based solutions for heart rate monitoring.

2.3.1 Wearables for heart rate monitoring

Heart rate is typically measured using EEG, ECG, or PPG sensors. However, EEG is limited to clinical settings, and ECG measurements during movement require a chest strap, making it impractical for everyday use. These limitations hamper the real-world applicability of these methods. While some smartwatches, such as the Apple Watch Series 9, can capture ECG readings, they require the user to close the circuit by touching the device with their fingers. Additionally, they only work when the user is stationary, making them unsuitable for continuous, in-the-wild monitoring.

To overcome these limitations, PPG has become the standard for heart rate monitoring in wearables. However, PPG is also limited by its susceptibility to motion artefacts caused by physical activity or body motion [51]. Bent et al. [52] found that among consumer and

research grade wrist-worn wearables, heart rate estimation errors increased by 30% during activity compared to rest. One specific problem with PPG is the signal crossover effect, where the sensors mistakenly lock onto periodic signals generated by motion, such as walking or running, instead of the heart signal [52,53], leading to inaccurate measurements. Moreover, PPG performance has been shown to vary based on skin tone, with higher measurement errors observed in participants with darker skin tones [52, 54]. These limitations highlight the need to explore alternatives to PPG sensors for heart rate monitoring.

Acoustic sensors on wearable devices have also been studied as a method for heart rate measurement. Chen et al. [55] estimated heart rate using a small acoustic sensor placed on the neck, and Sharma et al. [56] captured heart sounds with a microphone positioned on the radial artery at the wrist. Rose et al. [57] showed that cardiac sounds could be captured using microphones placed on the forehead, wrist, or ankle. These studies prove the feasibility of capturing cardiac-related sounds from various distant body locations when the user is stationary. However, these approaches are impractical for daily use due to the need for specialised, non-commodity wearable devices.

Therefore, there is still a need for a heart rate sensing modality which leverages commodity wearables and can provide accurate measurements during full-body motion and active scenarios.

2.3.2 Earables for heart rate monitoring

Research has also explored the feasibility of heart rate monitoring using earables, primarily using PPG and microphones.

Le Boeuf et al. [28] used a pair of PPG-embedded earbuds developed by Valencell Inc.[1] to determine heart rate, energy expenditure and VO_2max. Their study found good agreement between ECG and PPG-derived heart rates. Ferlini et al. [58] developed a custom PPG earbud and evaluated its performance across various activities. They reported errors of 27.14%, 29.84% and 12.52% for walking, running and talking respectively, quantitatively demonstrating that the signal crossover effects present in watch-based PPG also exist in ear-based PPG. This emphasises the challenges of PPG for heart rate estimation under motion.

Goverdovsky et al. [1] were the first to demonstrate the feasibility of capturing heart sounds

[1] https://valencell.com/

using passive sensing with an in-ear microphone, leveraging the occlusion effect. They showed that clear cardiac sounds could be captured while stationary as long as there was no jaw motion. Martin and Voix [10] extended this work by capturing heart sounds and computing heart rate while participants were stationary, using simple signal processing-based approaches without accounting for motion resilience. However, they found that even minor body movements while stationary introduced significant artefacts in the signals. Nirjon et al. [59] introduced an earphone equipped with an in-ear microphone to measure heart rate and an IMU to assess activity levels. However, they did not investigate the impact of activity on in-ear heart sounds or evaluate performance during motion. Gilliam et al. [60] used a custom in-ear microphone to determine heart rate and inter-beat intervals in stationary participants using simple signal processing. They also demonstrated the possibility of predicting atrial fibrillation from in-ear audio signals. Fan et al. [61] built a custom PCB to convert the speakers in commodity headphones to an in-ear sensor. Using this, they measured heart rate while stationary and while putting the earphones on and off by determining periodicity using autocorrelation. However, they did not examine the effects of full-body motion. Cao et al. [44] developed a sensing platform which provides access to the in-ear microphone on commodity ANC earphones. They demonstrated that with access to raw microphone signals from these devices, it is possible to obtain clear heart sounds while stationary, thus validating the potential of using commodity earables for heart rate monitoring. However, their approach relied on simple envelope and threshold-based methods, which are only effective on noise-free signals and when the user is stationary. Therefore, while numerous studies have examined the potential of in-ear audio for heart signal monitoring in stationary conditions, literature consistently finds that motion, even minor non-full-body motion, introduces significant artefacts that obscure heart sounds. No solutions have been proposed to reduce the impact of motion, nor has any work extended this monitoring to the more challenging scenario of full-body motion. Thus, accurate heart rate monitoring under motion using the sensors in commodity earbuds remains an unsolved yet critical problem.

2.4 Respiratory Rate Monitoring

Respiratory rate provides crucial information about overall health and fitness within the human body. Clinically, it is used for disease diagnosis and management of various conditions, and it can also serve as an early warning sign of health deterioration, such as in cases

of cardiac arrest or respiratory illnesses [62]. In daily life, respiratory rate indicates the presence of physical and mental stressors, including emotional stress, emotional response, and cognitive load [62]. Additionally, it is a key indicator of exertion levels during physical activities [63], offering valuable insights for optimising workout routines and detecting exercise-induced fatigue [64]. Therefore, continuous respiratory rate monitoring across various daily activities and settings is vital for gaining meaningful insights into health and fitness. In this section, we review existing research on respiratory rate monitoring using mobile and wearable devices and outline the gaps and limitations in current approaches. Then we explore the potential of breathing-rate monitoring using earables.

2.4.1 Mobile and wearable devices for respiratory rate monitoring

Respiratory rate monitoring has been extensively explored within the mobile computing community, but a comprehensive solution to this problem has yet to be developed.

Under motion conditions, respiratory rate is most commonly measured using a chest strap, similar to heart rate monitoring. The Zephyr [65] system uses pressure sensors in a chest strap to capture breathing movements and thus determine respiratory rate. However, these straps are uncomfortable and inconvenient for everyday use.

Research has also explored the use of smartphones for respiratory rate monitoring [11,66–70]. Most commonly, these studies utilise the smartphone's IMUs, though these solutions require the user to hold the device in a specific position. For instance, some approaches necessitate placing the smartphone on the chest [68–70], while others require it to be held against the abdomen [69]. Nam et al. [67] used the smartphone camera to capture chest movements, thereby determining respiratory rate, while another study by Nam et al. [11] recorded breathing sounds using the smartphone's built-in or headset microphones placed near the suprasternal notch or nose. Wang et al. [66] employed active acoustic sensing to monitor breathing-induced chest movements, but this method also requires the smartphone to be held in a specific posture. All these solutions, however, only work under stationary conditions and demand active user involvement with precise device placement, making them unsuitable for continuous or real-life monitoring.

On smartwatches, systems typically use either IMU or PPG sensors for respiratory rate monitoring. Several studies [71–74] have employed IMUs, though they focus on stationary conditions, such as rest, meditation, or lying down [71,72,74]. Liaqat et al. [73] demonstrated the possibility of estimating respiratory rate under both stationary and walking conditions

using learning-based techniques. However, this approach rejects sensor data with excessive noise, leading to only 20% of data windows being accepted during walking. As a result, accurate results rely heavily on significant data rejection.

For PPG-based monitoring, some commercial smartwatches [75, 76] have integrated respiratory rate estimation functionality, however, these only work at rest, such as during breathing exercises, while performing yoga, or when sleeping. Various studies have also examined PPG-based respiratory rate estimation on smartwatches [77, 78]. Zhao et al. [77] developed a learning-based approach for both sedentary and moving conditions, but the system struggles to provide reliable estimations during motion with unacceptably large errors. Dai et al. [78] focused on scenarios where users engage in discontinuous activities while sitting, such as watching videos, talking, doing math, or holding an object, but did not address more dynamic conditions. Therefore, smartwatch-based solutions for respiratory rate monitoring have shown promise using learning techniques. However, their reliability is compromised by motion artefacts, since on-the-wrist PPG and IMUs are both highly susceptible to motion-induced disturbances [74, 79, 80]. This susceptibility leads to low data retention rates or high estimation errors, particularly during movement.

Other approaches include using wireless signals like RF [81, 82] and WiFi [83], which have achieved considerable success under stationary conditions. However, these methods often cannot operate effectively in the presence of multiple users, outside the home, or under motion conditions. Additionally, some studies have employed body-worn PPG and ECG sensors [84–88] to estimate respiratory rate indirectly from heart signals, however, these also only function when the user remains still.

In summary, although there is extensive research on respiratory rate estimation using wearable devices, no solutions provide comprehensive monitoring across a range of daily activities and conditions.

2.4.2 Earables for respiratory rate monitoring

Research into respiratory rate monitoring using earables has explored various approaches and sensing modalities, including PPG, IMUs, combinations of sensors, and both out-ear and in-ear microphones.

Several studies have used in-ear PPG for respiratory rate monitoring [89–91], but these solutions are only effective under stationary conditions. Additionally, Taniguchi and

Nishikawa [89] required users to control their breathing at specific rates for optimal results. Only Ferlini et al. [4] investigated in-ear PPG for respiratory rate estimation across various user activities, such as stationary, talking, walking, and running. However, their findings showed that physical activity significantly affected accuracy, with error rates reaching approximately 31% under motion. Furthermore, since PPG sensors are not commonly integrated into commercial earables, the real-world applicability of these systems is limited.

Preliminary works [92–94] have also explored the use of IMUs on earphones to estimate respiratory rate. However, breathing-induced motions are very weak and easily overwhelmed by head or body movements. As such, these methods are only effective under stationary conditions by discarding data from periods with motions. These systems cannot function in the presence of full-body motion, such as during walking or running, where breathing motions are entirely overwhelmed.

Kumar et al. [95] estimated respiratory rate using out-ear microphones on AirPods, employing deep learning techniques. This approach relies on audible breathing sounds, which are retained through perceptual annotation for model training and testing, which means it only works effectively for heavy breathing. Moreover, out-ear microphones are inherently vulnerable to environmental noises since breathing sounds are weak and attenuate significantly in air. Ahmed et al [96] also employed out-ear microphones and IMUs on earphones for respiratory rate estimation, but their system was tailored for stationary conditions with head movement. However, in practical settings, natural breathing tends to be weak, making these approaches less effective. Some studies, such as [97, 98], have utilised out-ear microphones on earphones and IMUs on smartphones to estimate locomotor respiratory coupling (LRC) during running. This coupling can be used to estimate respiratory rate, but it only works during running and these works assume a constant LRC ratio, which is unrealistic in daily life scenarios.

Pressler et al. [99] investigated the characteristics of breathing sounds captured inside the occluded ear canal, and found that breaths above some "critical flow" rate (i.e. minimum quantity of air during a breath) create sounds that can be captured inside the ear. Goverdovsky et al. [1] also demonstrated the feasibility of capturing breathing sounds within the ear canal. They hypothesised that the sounds of turbulent airflow created by breathing travel through the tissues of the head and the eustachian tube to the ear. They showed that these can be captured, however concluded that they are weak and affected by motion artefacts. Martin and Voix [10] further explored this by estimating respiratory rate from

sounds captured with an in-ear microphone inside the occluded ear canal. They achieved this only when stationary and only using high-intensity (deep) breathing. Therefore, it has been shown that when the user is stationary, breathing sounds above a certain intensity (*i.e.,* deep breathing) can be captured using the in-ear microphone.

Thus, while there has been considerable research into respiratory rate estimation using earables, no system has been developed that is robust to movement and suitable for daily life activities. Therefore, the significant challenge of accurate respiratory rate monitoring under real-life conditions remains unsolved.

2.5 Stroke Volume Monitoring

2.5.1 Stroke volume primer

Stroke volume is a crucial cardiovascular parameter that plays a fundamental role in understanding how the heart meets the body's metabolic demands [13]. It refers to the volume of blood ejected by each ventricle with each contraction (each heartbeat) [100]. Stroke volume is closely related to cardiac output, another essential cardiovascular parameter that indicates the total volume of blood pumped by the heart per minute [100]. Cardiac output is a key indicator of the body's ability to deliver oxygen to the tissues and reflects ventricular function. The relationship between stroke volume (SV) and cardiac output (CO) is given by Equation (2.1), where HR represents heart rate [100].

$$CO = HR * SV \tag{2.1}$$

To fully understand stroke volume, it is important to first explain the cardiac cycle, as illustrated in Figure 2.5.1 [100, 101]. The systolic phase of the cardiac cycle begins with the QRS complex on the ECG, which indicates ventricular depolarisation. This electrical event triggers ventricular contraction and produces the S1 heart sound, which is produced by the closure of the valves between the atria and ventricles as blood is ejected from the heart. This is followed by the T wave on the ECG which indicates ventricular repolarisation.

After systole, the heart enters the diastolic phase. During diastole, the ventricles relax, causing ventricular pressure to fall below that of the aorta and pulmonary artery. This drop in pressure leads to the closure of the aortic and pulmonary valves, producing the

S2 heart sound. As the ventricles continue to relax, they begin to fill with blood from the atria, increasing ventricular volume.

At the end of diastole, the ventricle reaches its maximum blood volume, known as the end-diastolic volume (EDV). After the ventricular contraction during systole, some blood remains in the ventricle, referred to as the end-systolic volume (ESV). The difference between the EDV and ESV (EDV - ESV) represents the stroke volume — the amount of blood ejected from the ventricle during each heartbeat.

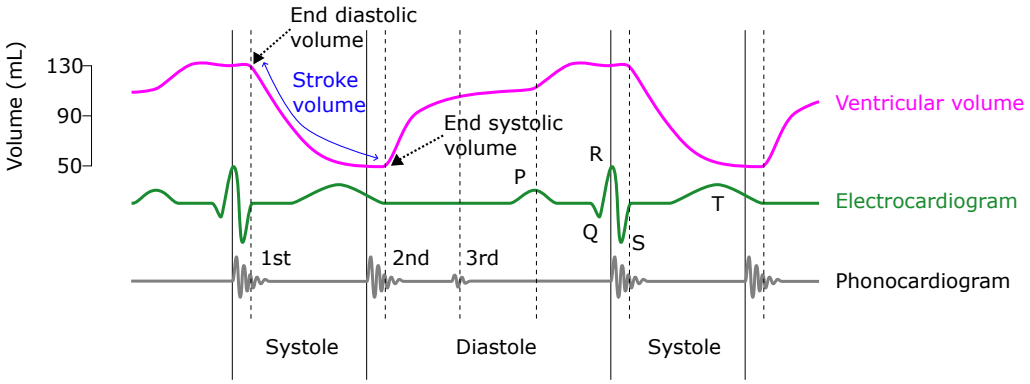


Figure 2.5.1: Phases of the cardiac cycle illustrated using heart sounds, electrocardiogram signals and ventricular volume. Adapted from [102].

Stroke volume is influenced by multiple physiological factors that affect the efficiency of the heart [100, 103]. These are as follows:

- Preload: The amount of stretch in the heart muscle fibers right before contraction. The preload is related to the EDV and proportional to the stroke volume.

- Afterload: The arterial pressure (or resistance) the ventricles must overcome to eject blood during systole. Afterload is determined by arterial blood pressure and systemic vascular resistance and is inversely proportional to stroke volume. An increased afterload, such as in hypertension, can reduce stroke volume.

- Contractility: The strength of the heart muscle's contraction, which is influenced by the magnitude of sympathetic nervous system input to the ventricles.

2.5.2 Clinical measurement of stroke volume

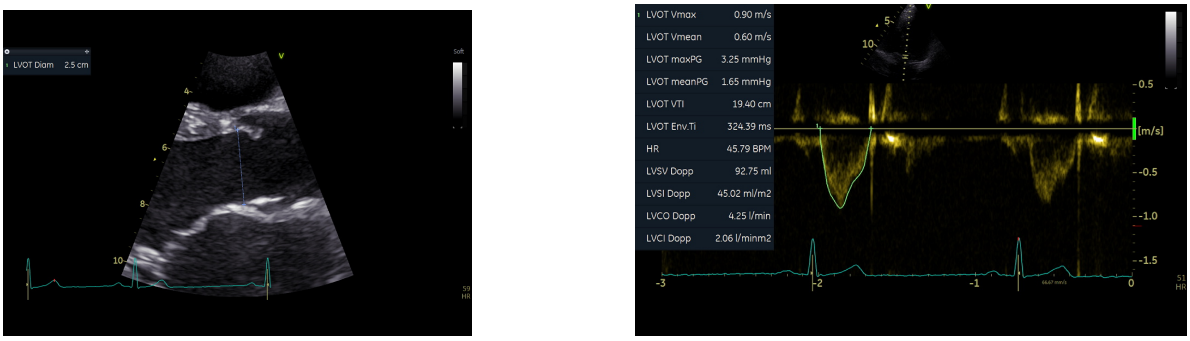
The gold standard for measuring stroke volume is the Direct Fick method, which involves using pulmonary arterial or transpulmonary catheters [104, 105]. This method begins with calculating the total oxygen consumption by comparing the oxygen content in the subject's inspired and expired breaths [105]. Catheters are then used to obtain venous blood samples from the pulmonary artery and arterial blood samples from any other artery. The difference in oxygen content between the arterial and venous blood is computed, and cardiac output is determined by dividing the total oxygen consumption by the arterial-venous oxygen difference [105]. Finally, stroke volume is calculated by dividing cardiac output by heart rate. However, this method is highly invasive, as it requires multiple catheters, and labor-intensive [106].

As a less invasive alternative, echocardiography (or cardiac ultrasound) has become increasingly popular for measuring stroke volume [105]. Echocardiography uses ultrasound waves to visualise the heart and assess blood flow. Stroke volume can be determined either by measuring the end-diastolic volume (EDV) and end-systolic volume (ESV) or by using the Doppler method, as shown in Equation (2.2). In this equation, VTI represents the velocity time integral of the Doppler flow profile of the blood, and CSA is the cross-sectional area of the left ventricular outflow tract, as depicted in Figure 2.5.2(a).

$$SV = VTI * CSA \tag{2.2}$$

Figure 2.5.2(b) shows an echocardiogram with the Doppler flow profile, where the VTI is represented by the blue traced curve, and the resulting stroke volume is labelled as *LVSV Dopp*. Despite being non-invasive, echocardiography requires highly skilled operators and expensive equipment to accurately determine stroke volume, limiting its widespread use [105].

For non-invasive and continuous stroke volume measurements, finger-cuff-based systems, such as the Finapres NOVA [107], can be used. These systems use an inflatable finger cuff containing PPG sensors to estimate arterial blood pressure and stroke volume. The finger cuff employs the volume-clamp method for beat-to-beat blood pressure estimation [108]. By maintaining a constant diameter of the finger artery, any changes in cuff pressure directly reflect changes in blood pressure [108]. The aortic flow waveform is then derived from the arterial blood pressure, and stroke volume is estimated as the integral of the flow waveform



(a) Measurement of the LVOT.

(b) Measuring stroke volume by computing the VTI of the blood flow through the heart.

Figure 2.5.2: Echocardiogram for stroke volume measurements.

per beat. Although these methods are reliable, they require accurate calibration with stroke volume measurements obtained from an echocardiogram [109].

Therefore, these methods only have clinical utility due to the complexity of the procedures involved, and the need for specialised, expensive equipment and expertise. Consequently, there is significant value in developing a new system for stroke volume measurement that is convenient, non-invasive, cost-effective, and suitable for use outside of clinical settings.

2.5.3 Non-wearable based methods for stroke volume measurement

To overcome the limitations associated with invasive procedures, specialised skills and expensive equipment, research has explored non-wearable based methods for stroke volume measurement which rely on physiological proxies for stroke volume.

Systolic time intervals have been used clinically as proxies for stroke volume [109]. One key interval is the left ventricular ejection time (LVET), which is the duration between the opening and closing of the aortic valve (*i.e.,* the period during which the ventricle is ejecting blood) [110]. Under normal cardiac function, stroke volume and LVET are proportional, where a greater volume of blood ejected by the ventricle results in a longer ejection time [109]. LVET is also related to contractility and afterload, similar to stroke volume. Finkelstein et al. proposed a linear model for estimating stroke volume that includes LVET, as shown in Equation (2.3), where BSA is body surface area [109].

$$SV = -6.6 + 0.25 * LVET + 40.4 * BSA - 0.51 * Age - 0.62 * HR \tag{2.3}$$

Changes in afterload and the resulting pressure variations also impact the pre-ejection period (PEP), which is the time between the Q wave of the ECG and the opening of the aortic valve [110]. PEP is therefore also correlated with stroke volume. Couceiro et al. [109] expanded on the Finkelstein model by incorporating PEP and the PEP/LVET ratio into a non-linear model using a feed-forward neural network. Their study demonstrated that using ground truth timing metrics obtained using echocardiograms, the non-linear model significantly improved stroke volume estimations compared to the Finkelstein model. Even when the timing metrics were derived from heart sounds obtained through auscultation, the model still achieved superior performance, achieving a Pearson correlation of 0.77 and an average error of 10%. This highlights the importance of timing-related metrics derived from heart sounds for accurate stroke volume estimation.

In addition, Shin et al. [111] demonstrated a correlation between the amplitude of the first heart sound, measured using a stethoscope, and cardiac output. This finding suggests, through Equation (2.1), that the amplitude of the first heart sound is proportional to stroke volume.

Cardiac output can also be modelled using the two-element Windkessel model shown in Figure 2.5.3 [15, 112]. In this model, cardiac output is calculated as the mean arterial blood pressure divided by the total peripheral resistance. Wang et al [104] demonstrated that proxies for arterial blood pressure and resistance can be derived by extracting features from PPG signals to estimate cardiac output. Lee et al. [113] developed a multivariate regression model to estimate cardiac output using spectral and morphological features extracted from finger-based PPG. These features include pulse width (a measure of PPG amplitude) and spectral power across low, mid, and high-frequency ranges. Therefore, cardiac output, and by extension stroke volume, is also related to various features of the PPG signal.

Yazdi et al. [104] evaluated the effectiveness of a novel scale equipped with ECG, ballistocardiogram and PPG signals for estimating stroke volume and cardiac output. Using features related to pulse pressure and timing-related metrics (PEP and LVET), they achieved an overall Pearson correlation of 0.81 with a percentage error of 36.73%. However, the scale requires specialised sensors, making it expensive, and it relies on user adherence which is a well-known challenge with medical devices [114].

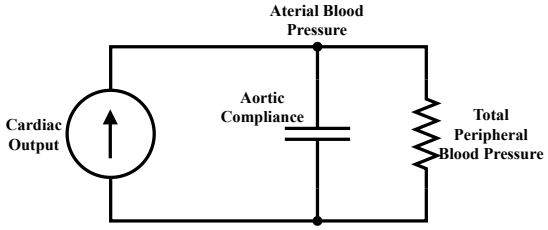


Figure 2.5.3: Two-element Windkessel model of cardiac output. Adapted from [15].

Therefore, previous work has leveraged physiological proxies for cardiac output and stroke volume estimation using various cardiac-related signals combined with processing techniques. However, these approaches have been limited to clinical settings and are therefore inapplicable for daily use. Despite this, they demonstrate the feasibility of determining stroke volume using various cardiac signal modalities.

2.5.4 Wearable devices for stroke volume measurement

To improve the accessibility of stroke volume measurements outside of clinical settings, limited recent studies have focused on developing wearable-based systems for stroke volume estimation.

Ganti et al. [13] designed a novel chest-worn sensor equipped with a single-lead ECG and a triaxial accelerometer to capture seismocardiogram signals (heartbeat-induced chest vibrations). Using this device to measure stroke volume, they achieved an R^2 score of 0.76 with a root mean square error of 11.48 mL using a regression model with features related to LVET and PEP. Wang et al. [15] employed PPG signals collected from the finger to estimate cardiac output, achieving a percentage error of 16.2% by using a novel index derived from the frequency domain of the PPG signal related to pulse pressure.

Dvir et al. [14] developed a custom PPG chest patch to determine cardiac output. However, this patch required calibration from a blood pressure cuff for accurate measurements. Nachman et al. [115] validated cardiac output measurements using a PPG device attached to a pig's tongue, however, this device also required calibration.

Despite these advancements, these devices rely on specialised hardware, often with multiple specialised sensors, making them expensive, cumbersome to wear and less socially acceptable for daily use. Currently, no solution exists that uses only commodity hardware for in-the-wild stroke-volume monitoring.

2.6 Audio Processing Techniques

Audio processing involves manipulating and analysing audio signals to enhance them and extract meaningful features. Audio processing has its foundation in signal processing, with modern advances predominantly coming from deep learning. One of the key challenges in earable research for physiological monitoring is the presence of noise in bio-signals. Effective techniques are therefore needed to mitigate the effect of noise and extract cleaner signals for processing. Many of the techniques discussed in the following section focus on this task. Another challenge is developing algorithms that are generalisable to the diverse signal properties across different users. To address this, we detail deep learning-based techniques for both signal denoising and predicting physiological parameters.

2.6.1 Audio signal-processing

In this section, we describe several commonly used signal-processing techniques for audio which are leveraged in this thesis.

Fast Fourier Transform and Short Time Fourier Transform

The Fast Fourier Transform (FFT) and Short Time Fourier Transform (STFT) are techniques to examine the frequency content of a signal. The Fourier transform (FT) converts a time-domain signal into its frequency-domain representation. For digital signal processing, the FFT, a discretised, computationally efficient version of the FT, is used. However, the FT does not work for signals with frequency content which changes over time (*i.e.,* non-stationary signals).

The STFT addresses this limitation by providing both time and frequency information. With the SFTF, the signal is divided into short windows, and the FFT is applied to each segment, resulting in a time-frequency signal representation. However, there is an inherent tradeoff in the STFT between temporal and spectral resolution based on the size of the window, where it is not possible to simultaneously achieve high temporal and spectral resolution.

Despite these limitations, FFT and STFT are useful techniques in examining the main frequency content of a signal window, such as for estimating breathing rate or walking frequency.

Hilbert transform

The Hilbert transform is one of the most common techniques for envelope extraction. Envelope extraction is a key signal processing technique that extracts information about the modulation of a signal, focusing on the underlying signal shape without the influence of its extremes. For example, in an ECG signal containing heart activity, there exists a modulation with the respiratory frequency whereby the amplitude changes with varying respiratory rate. In a heart sound signal, the envelope outlines each heartbeat, removing the influence of higher frequency signal changes and enabling heart rate extraction, as illustrated in Figure 2.6.1. The Hilbert Transform is often used for heart sound segmentation [116] and heart rate extraction [117].

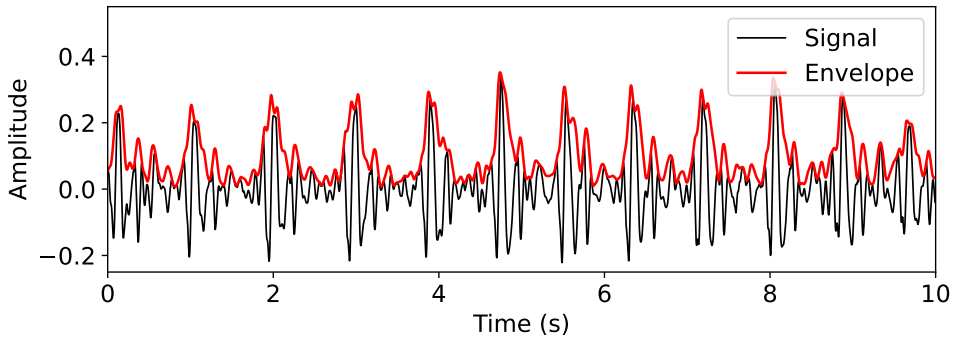


Figure 2.6.1: In-ear heart signal with its Hilbert envelope.

The Hibert transform is described by Equation (2.4), where $u(t)$ is the input signal and the Hilbert transform of $u(t)$ is computed by shifting the phase of each frequency component in $u(t)$ by $-\pi/2$. When added to the original signal, this phase shift creates an analytic signal. The magnitude of the analytic signal can then be extracted to obtain the signal envelope $E(t)$ as shown in Equation (2.5).

$$H(u(t)) = \lim_{\epsilon \to 0} \frac{1}{\pi} \int_{|s-t|>\epsilon} \frac{u(s)}{t-s} ds \tag{2.4}$$

$$E(t) = \sqrt{H(u(t))^2 + u(t)^2} \tag{2.5}$$

Discrete wavelet transform

Wavelet transforms are linear signal transforms, which are specifically suited for analysing signals whose frequency content changes over time [116]. The wavelet transform decomposes the signal into wavelets (or windows) at different scales and positions in time, providing information about how the frequency content of the signal changes with time [116]. The finite-length wavelets used in the wavelet transform make it better than the FFT at identifying localised features in a signal.

The discrete wavelet transform (DWT) is a computationally efficient version of the wavelet transform used in digital signal processing. The DWT works by applying pairs of filters to the signal: a high-pass filter to capture high-frequency components and a low-pass filter to capture low-frequency components [118, 119]. The high pass filters are related to the chosen wavelet function, and optimal results are obtained by selecting a wavelet function with characteristics similar to those of the signal. In the DWT, multiple pairs of filters are applied to decompose the signal into multiple levels, extracting information across different frequency ranges [118, 119]. Due to its feature localisation ability, the DWT is well-suited for signal denoising and has been frequently used for chest-based heart sound denoising [116]. When using DWT-based filtering, thresholds are applied to the DWT coefficients to remove small coefficients (likely representing noise), thus reducing the noise content of the reconstructed signal [119]. However, DWT-based filtering is dependent on selecting the correct wavelet and may fail if the frequency content of the noise cannot be separated from that of the signal itself.

Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) is a technique used to decompose a time series signal into several independent components including trends, periodic components, and noise [120]. Once the signal has been decomposed, the relevant components can be recombined to reconstruct the series of interest. SSA can be used to remove noise from a signal, or to select the underlying oscillatory components from a signal, *e.g.,* identifying the oscillatory breathing signal within a signal containing breathing sounds, heart sounds and footstep sounds. SSA has been successfully used for removing ECG interference from EMG signals [120], removing heart sounds from breathing signals [121] and reducing the impact of head motion on IMU signals of human gait [122].

We use the above-mentioned signal processing techniques for audio in Chapters 3 and 4.

2.6.2 Deep learning techniques for audio processing

The advent of deep learning has introduced new techniques and tools for processing audio signals. Specifically, by converting audio windows into spectrogram representations using the STFT, we can apply deep learning architectures originally developed for images, such as convolutional neural networks (CNNs) [123]. Of particular interest are autoencoders and particularly the models that build upon the principle of the autoencoder. Autoencoders are networks designed to reconstruct the input signal at the output by learning efficient representations of unlabelled data using an encoder-decoder architecture. The encoder compresses the input to a lower-dimensional representation thereby learning the most important features from the input data while removing irrelevant ones.

The decoder then reconstructs the input data from these features. Training aims to minimise the error between the reconstructed output and the original input. Autoencoders can also be used for denoising by adding synthetic noise to the input signal and minimising the loss between the clean input and the reconstructed output [124]. Autoencoders have been successfully used in both the image and audio domains, particularly in speech enhancement and denoising [125].

U-Net

U-Net is a specific autoencoder architecture that leverages CNNs, making it good at capturing both local and global signal features [126]. U-Net was initially developed for medical image segmentation, however, it has proven effective in audio processing, such as speech enhancement [127], and for image denoising [128]. For denoising tasks, U-Net is trained on a dataset of paired noisy and clean samples. The network learns to map the noisy inputs to their corresponding clean outputs thereby removing the noise. A challenge in physiological signal processing lies in obtaining noise-free samples that can be used for denoising. We address this challenge in Chapter 3.

The U-Net architecture consists of a symmetrical encoder-decoder structure built upon convolutional layers:

- **Encoder**: The encoder, or contracting path, consists of multiple repeated blocks, each containing a 3×3 convolutional layer followed by max pooling to downsample the data. With each block, the data is downsampled further, facilitating the extraction of higher-level features.

- **Decoder**: The decoder, or expansive path, consists of blocks for upsampling the feature map, applying a 2×2 convolution, and concatenating the result with the corresponding feature map from the encoder, followed by a 3×3 convolutional layer with max pooling. The upsampling layers increase the dimensions of the feature maps to restore them to their original resolution.
- **Skip Connections**: The concatenation of the feature maps from the encoder and decoder is achieved through skip connections. These give the decoder access to the higher-resolution features from the earlier layers, allowing the network to recover fine-grained details that were lost during downsampling.

This thesis uses the U-Net architecture in Chapter 3.

Vision Transformer

Vision transformers are based on the transformer architecture, which was initially developed for natural language processing (NLP). In vision transformers, images are treated as a sequence of patches, similar to a sequence of words in an NLP task [129]. Transformers employ self-attention mechanisms to capture dependencies across the full input sequence, enabling the model to understand long-range dependencies [129]. Self-attention allows the model to focus on different parts of the input sequence when making predictions, thus capturing complex patterns and temporal dependencies, which is critical for physiological signals [129]. Additionally, transformers use multi-head attention which enables the model to focus on different parts of the input sequence simultaneously. This approach allows for the extraction of better data representations that account for temporal dependencies [129].

Within the context of this thesis, by converting an audio window to an image representation (normally a spectrogram), image-based transformers can be applied to audio.

The architecture is as follows [129]:

- **Patch embedding:** The spectrogram is divided into patches and flattened into a 1-dimensional vector. Positional embeddings are added to the patches to retain their order.
- **Encoder:** N transformers [129] are stacked together to form an encoder. Each transformer is composed of a multi-head attention layer and a feed-forward network.
- **Decoder:** N transformers are also stacked together to form a decoder. These have

the same layers as the encoder transformers, with an additional layer that performs multi-head attention over the output of the encoder [129].

- **Linear head:** At the top of the decoder, a linear head reconstructs the input spectrogram.

This vision transformer forms the backbone of the masked autoencoder used in Chapter 5. A masked autoencoder uses a vision transformer to predict masked pixels from an input image [130]. It is an asymmetric autoencoder, meaning the encoder learns only from the unmasked pixels (reducing computational complexity), while the decoder reconstructs the full input signal [130]. Once the masked autoencoder has been trained, the decoder can be removed, and the encoder used for downstream tasks.

2.6.3 Deep learning training paradigms

In addition to the neural network architectures described in the previous section, different training techniques can be leveraged to improve performance. A key limitation of deep learning techniques is their need for large quantities of data [131]. As discussed in Chapter 1, data scarcity is a key challenge in earable and wearable research. Thus, it is necessary to use training paradigms that can overcome the mismatch between the high data requirements of deep learning and the low data availability of earable studies. To achieve this, in this thesis, we use transfer learning and self-supervised learning.

Transfer learning

Transfer learning is a training technique in which a model trained on one task is reused for another, related task [131]. This approach enables the model to leverage its previous knowledge to improve performance on the target task. Transfer learning is built on the human experience where we apply prior knowledge to solve new problems better and faster [131]. Transfer learning involves training a baseline model on a large dataset and then fine-tuning its weights (*i.e.,* learned representations) using the target dataset to achieve the required task (the downstream task). The initial learning can be either supervised or self-supervised (as discussed in the following section). Transfer learning is a powerful tool for overcoming data scarcity, reducing model overfitting and improving generalisability.

This thesis uses transfer learning in Chapters 3 and 5.

Self-supervised learning

Traditionally, neural networks were trained using supervised learning, where the models learn from labelled data, mapping the input to its associated output. However, supervised learning is not always feasible in situations where labelled data is scarce or unavailable. In such cases, it can be advantageous for a model to learn from unlabelled data, which is often easier to obtain. For example, it is challenging to collect a dataset of PPG data with corresponding ground truth labels from an ECG chest strap. However, due to the large number of people worldwide wearing smartwatches, it would be easier to obtain a dataset of unlabeled PPG data.

Self-supervised learning addresses this challenge by enabling models to learn from the data itself without requiring explicit labels [131, 132]. In self-supervised learning, the model learns to predict some aspect of the dataset, *e.g.,* the next frame in a video. One of the most common techniques of self-supervised learning is by masking, where the model applies binary masks to portions of the input and then predicts the content of the masked area [131]. This approach allows the model to supervise its own training by using the unmasked portion as the label, and the masked portion as the input [131]. Through this process, the model captures essential features from the data, which are crucial for reconstructing the input. Self-supervised learning is, therefore, a powerful technique for improving model robustness and generalisability [131].

This thesis uses self-supervised learning in Chapter 5.

2.7 Summary

This chapter has reviewed the state-of-the-art in physiological monitoring using earables and other wearable devices, with a focus on measuring heart rate, respiratory rate, and stroke volume. This chapter has also presented key signal processing and deep learning techniques which have been successful when applied to audio processing. While significant progress has been made in physiological monitoring, this review has revealed research gaps in the literature which motivates the designs of the systems proposed in this thesis.

In Section 2.3, we highlighted the need for a system that can accurately measure heart rate in the presence of motion and under various daily life conditions. This relies on the development of robust signal processing and machine learning techniques to accurately extract heart rate, even in the presence of motion artefacts. In Chapter 3, we propose a deep

learning-based signal denoising and enhancement algorithm to reduce the impact of motion artefacts, allowing for accurate heart rate estimation from in-ear audio. Additionally, we explore the use of transfer learning to address the limitations posed by small datasets in deep learning tasks. Our work demonstrates the feasibility of motion-resilient heart rate monitoring using commodity sensors.

Section 2.4 showed that while respiratory rate monitoring has been extensively explored using wearable devices, most existing systems are limited to stationary conditions or require specific device placement, making them unsuitable for real-world monitoring. The few systems that work under motion conditions have large errors and low data retention. Respiratory rate monitoring systems also often struggle with poor SNR. This highlights a gap in the literature for a solution that can accurately and continuously measure respiratory rate under daily life conditions without requiring user intervention or specific device placement. To address this gap, in Chapter 4, we present a signal processing-based system that determines respiratory rate across various active and resting scenarios. To overcome the challenges posed by poor SNR, we leverage physiological couplings between respiratory rate and the cardiovascular system, as well as between respiratory rate and gait. We therefore present a new method for respiratory rate monitoring that overcomes the challenges and limitations found in the literature.

Next, we explored stroke volume, a key clinical metric for cardiovascular health in Section 2.5. This review proved that limited solutions exist for stroke volume measurement outside of clinical settings and without the need for invasive procedures, specialised equipment and expertise. To overcome this, in Chapter 5 we collect a novel dataset consisting of in-ear audio and corresponding stroke volume measurements and propose a system for measuring stroke volume using in-ear audio. This system does not require specialised equipment and uses the sensors on commodity earables, thus advancing the field of stroke volume monitoring in real-world settings.

Chapter 3

Motion-Resilient Heart Rate Monitoring

3.1 Introduction

In Chapter 1, we discussed the potential of earables for sensing human physiology, and the limitations of today's earables research. We also detailed the challenges facing in-ear sensing and heart rate monitoring using in-ear audio. In this chapter, we address these limitations by, for the first time, presenting a solution to heart rate monitoring under motion using in-ear audio on earables.

As discussed in Chapter 2, heart rate is one of the five main vital signs and is strongly associated with cardiovascular disease and mortality risk. It is also an excellent indicator of both physical and mental health. Monitoring heart rate, particularly under motion, is valuable for assessing cardiorespiratory fitness and ensuring that target heart rates are maintained during exercise. Additionally, by monitoring heart rate under motion, heart rate data can be used to tailor training programs and monitor recovery after exercise. Therefore, there is great value in monitoring heart rate both while sedentary and under motion.

As we showed in Chapter 2, while ECG remains the gold standard for heart rate monitoring, its form factor makes it cumbersome, uncomfortable, and inconvenient. In contrast, as described in Chapter 2, within wearables, PPG sensors, which measure light scatter as a result of blood flow, are the most common due to their non-invasiveness, easy implementation and low cost. Although PPG is effective and accurate for heart rate measurements

in stationary conditions [52], it is highly sensitive to motion artefacts caused by body movement or physical activities [52, 53, 133]. Due to these motion artefacts, the research community has yet to find an agreement on the goodness of wrist-worn PPG (e.g. PPG on smartwatch). While the topic has been widely investigated [52, 53, 133], a consensus on the best commercially available device to monitor the wearer's heart rate whenever motion is concerned, is yet to be found. Moreover, intense motion, like walking and running, yields substantial deviations from the ground truth heart rates, resulting in average errors up to 30% across a wide spectrum of wrist-worn devices [52]. Dealing with interference from motion artefacts is thus an open and challenging problem in heart rate estimation.

Given the limitations of wrist-based PPG, researchers have started investigating alternative wearables for heart rate monitoring under motion. With the rapid spreading of earables in daily life [2], they offer a portable and non-invasive means of continuous heart rate detection. Particularly, due to their pervasiveness during physical activity (specifically while walking and running), the earable form factor can be exploited for heart rate monitoring while under motion. Emerging research on earable-based PPG for continuous heart rate sensing [134] shows promise, but real-world performance under motion remains poor [53, 58]. In fact, similar to what is observed for wrist-worn devices [52], errors around 30% have been reported [58].

Recently, Martin and Voix [10] proposed to leverage in-ear microphones and the occlusion effect for heart rate estimation. Their results show an error of 5.6% for heart rate determination under stationary conditions. However, Martin and Voix [10] only demonstrated the feasibility of measuring heart rate with in-ear microphones while an individual is stationary: *how in-ear microphone heart rate measurement performs under active scenarios remains unclear and unexplored.*

In this chapter, to address this gap, we focus on in-ear heart rate estimation under both sedentary and active scenarios (i.e. walking, and running). The biggest hurdle to accurate heart rate measurement is motion-induced interference, which is amplified along with the heart sounds by the occlusion effect [42]. Removing such interference is non-trivial and poses two major challenges. First, the strength of heartbeats is much weaker than the foot strikes, so heartbeat signals are buried in the walking and running signals, and the heartbeats thus have a very low signal-to-noise ratio. Second, since heart rate and walking frequency (i.e. cadence) are similar (both around 1.5-2.3 Hz [135]), it is hard to separate them in the frequency domain.

To address these challenges, we propose a processing pipeline for accurate heart rate detection in the presence of motion artefacts, namely, walking and running. We leverage CNNs to develop a heart signal denoising and enhancement pipeline from which heart rate can be estimated. Different from previous audio-based heart rate estimation works [10,134,136,137], we also validate the functioning of our technique while music is played simultaneously into the ear, showing that the proposed approach can determine heart rate even when playing music through the earbud. With data collected from 15 subjects, we demonstrate that an in-ear microphone can be a viable sensor for heart rate estimation under motion cases with good performance. Specifically, with mean absolute percentage error (MAPE) of less than 10% while sedentary, walking and running, this system is accurate according to ANSI standards for heart rate accuracy for a physical monitoring device [52, 138]. Additionally, because of the artefacts considered, the vantage points (the ears), and the device form factor (earables), our work is directly comparable to Ferlini et al. [58]. Notably, we significantly outperform in-ear PPG [58] (71% and 68% improvement) for walking and running. This result hints at the great potential of in-ear microphones for cardiovascular health monitoring, even under challenging scenarios. Moreover, compared to PPG, microphones are more energy efficient [139, 140] and affordable offering additional appeal for continuous heart rate estimation.

The contribution of this chapter can be summarized as follows:

- We explore heart rate estimation with in-ear microphones and present an analysis of the interference imposed by full-body motion.
- We propose a novel pipeline for heart rate estimation under motion artefacts, consisting of a CNN-based module using the U-Net architecture to enhance audio-based heart sounds with ECG as a reference, and an estimation module using signal processing to estimate heart rate from cleaned signals. We further leverage transfer learning by pre-training the model using a large heart sounds dataset and fine-tuning it using our data to effectively capture heart sounds-related information, and handle the limited data size. To the best of our knowledge, no previous works have attempted to clean and enhance heart sounds captured by earables using ECG signals.
- We built a custom earbud prototype and collected data from 15 subjects. Results show that we can achieve mean absolute errors of 1.88 ± 2.89 BPM, 6.83 ± 5.05 BPM, and 13.19 ± 11.37 BPM for sedentary, walking, and running, respectively, demonstrating the effectiveness of the proposed approach in combating motion artefacts.

The techniques and results in this chapter have been published in [16] and [17].

3.2 Primer

In this section, we present the mechanism by which heart sounds are collected in the ear and the challenges of achieving accurate and portable in-ear heart rate estimation under motion conditions.

3.2.1 In-ear heart sound acquisition

As discussed in Section 2.2.1, by placing a microphone in the occluded ear canal, we can capture bone-conducted sounds. Heart sounds are among the sounds that are conducted through the bones to the ear canal. As such, we can capture heart sounds using earables. An example showing the heart sounds captured by the internal microphone is shown in Figure 3.2.1. The two sounds in the cardiac cycle (S1 and S2) can be captured using the in-ear microphone, thus indicating the potential of in-ear microphones for heart rate monitoring. The correlation between the in-ear captured audio and the ECG signal is also evident in Figure 3.2.1.

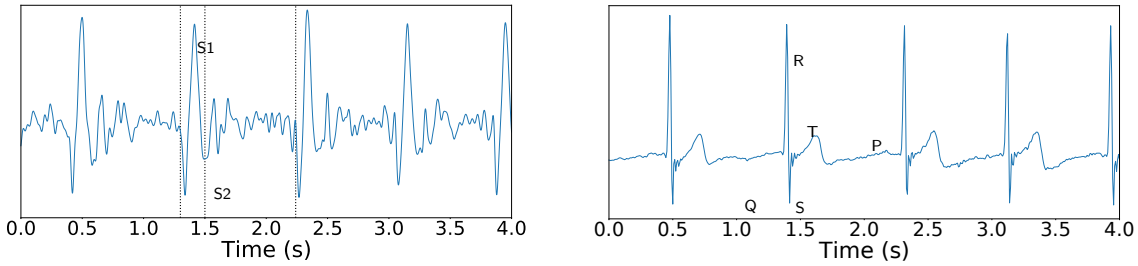


Figure 3.2.1: The (left) sound signal captured by the internal microphone indicating the S1 and S2 heart sounds, and the (right) corresponding ECG signal showing the three main components of the ECG signal (QRS complex (ventricular depolarization), T wave (ventricular repolarization), P wave (atrial depolarization)).

3.2.2 Motion artefacts analysis and challenges

In-ear microphone-based heart rate estimation suffers from human motion artefacts since the occlusion effect not only amplifies the heartbeat-induced sound, but also enhances other bone-conducted sounds and vibrations inside the body [42, 45]. Figure 3.2.2(a-c) illustrates the recorded audio signals from the in-ear microphone while stationary, walking

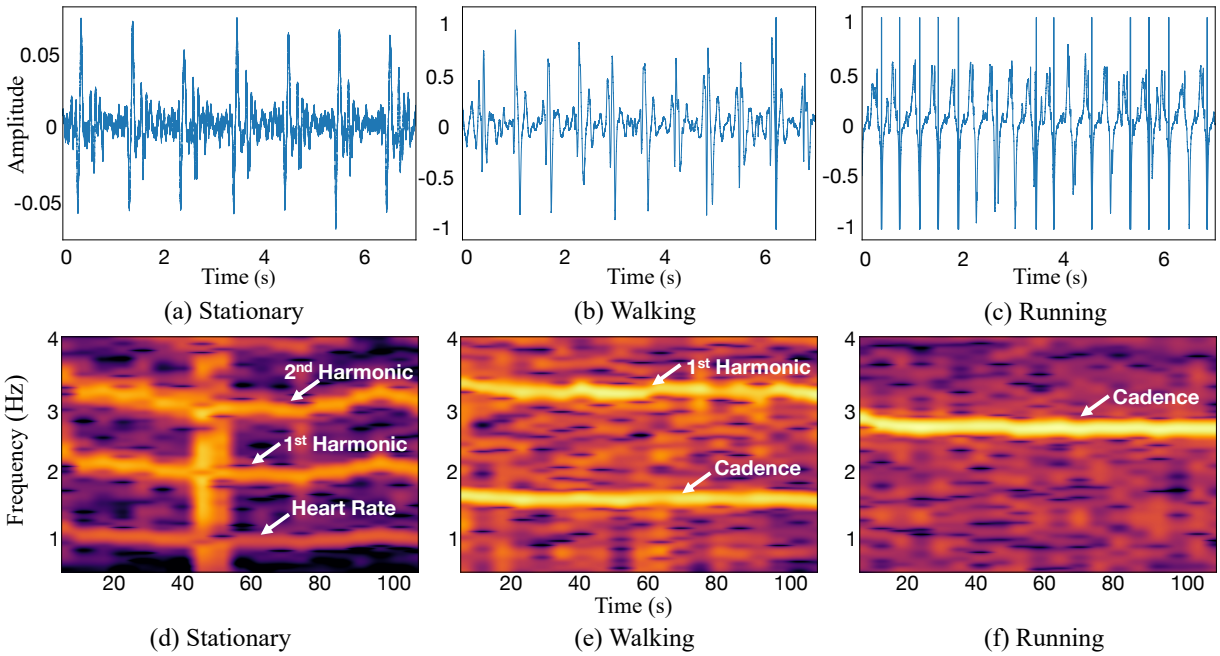


Figure 3.2.2: Time domain representations and spectrograms of audio signals captured by the in-ear microphone.

and running within a seven-second window. Figure 3.2.2(d-e) are spectrograms of the activities shown over a longer timescale so that trends in heart rate can be seen. The heartbeat is clearly observable when an individual is stationary in Figure 3.2.2(a), with heart rate frequency lying around 1 Hz and its 1st and 2nd harmonics clearly observable in Figure 3.2.2(d). In contrast, the heart signal is completely overwhelmed by the amplified step sounds in Figure 3.2.2(b) (note the different scales of the y-axis), with the periodic peaks corresponding to the sound of foot strikes propagating through the human skeleton, resulting in significantly higher energy observed around 1.7 Hz (the cadence) in Figure 3.2.2(e). Here, we can no longer see the heart rate or its harmonics, as they are overwhelmed by the walking cadence. Additionally, the frequency content of the heart sounds and the step sounds overlap, making it difficult to separate the two signals. This overlap presents a challenge for accurately estimating the heart rate, both in the time and frequency domain.

The heartbeats are further affected and obscured by foot strikes while running (Figure 3.2.2(c)) which exhibit far stronger energy than any of the other activities, with high energy at 2.8 Hz (Figure 3.2.2(f)) again corresponding to the cadence. Here the low-amplitude heart sounds are entirely buried in the high-amplitude step sounds.

3.3 System Design

Typical signal processing techniques have shown effectiveness in heart rate estimation in the stationary case [10]. However, they do not adequately isolate the heart sounds from the corrupted audio under motion artefacts, since motion-artefact elimination is a non-trivial problem. Typical signal processing techniques for denoising are more effective under certain signal-to-noise ratios (SNR) and errors increase with decreasing SNR [141,142]. Additionally, the differences in the user's anatomy (different ear canal shapes, different earbud fit levels and thus changes in the resultant amplification) result in differences in the captured audio sounds, and this variability is poorly captured and processed using signal processing. Due to the recent successes witnessed by deep learning for denoising in numerous fields [143,144], we propose a novel pipeline using deep learning to eliminate motion artefacts in audio and estimate heart rate. In the following sections, we first present a signal processing approach for heart rate estimation and then the proposed deep learning pipeline for motion artefact removal.

3.3.1 Signal processing for heart rate estimation

The initial phase of our work involved the development of a signal-processing pipeline for heart rate estimation. This aims to provide an efficient and computationally effective heart rate detection method, and to explore the potential of typical signal processing techniques in heart rate estimation under motion artefacts.

First, we compute the Hilbert transform (introduced in Chapter 2) of the audio to calculate the heart rate envelope. We then compute the spectrum of the envelope using the FFT and detect the dominant peaks which are converted to the heart rate. This approach performs well on a clean and stationary signal (see Section 3.5.2). However, when audio signals are corrupted with motion, dominant peaks in the spectrum correspond to motions, rather than the heart rate, thus introducing errors in heart rate estimation. More sophisticated denoising techniques are thus required to obtain clean heart sounds under motion. The DWT (described in Chapter 2) is therefore used to remove artefacts from the audio to isolate heart sounds. Specifically, we filter out detail coefficients from the DWT based on signal variance, thus removing the noise components with a high variance from the mean.

Though denoising can yield a relatively clean heart sound signal, the denoised signals are still interfered with by the motion artefacts to some extent, due to the underlying

complexity of the artefacts, and the closely overlapping frequency ranges of the artefacts and the heart sounds. Therefore, we propose a frequency spectrum searching algorithm using the FFT (as introduced in Chapter 2) to estimate the HR from the denoised signal to account for the remaining motion artefacts. Instead of searching the FFT peaks over the full frequency range of the denoised audio, we only search the heart rate peaks in a small frequency range corresponding to the range of allowable human heart rates and the heart rate in the previous window. This guarantees that the peaks in heart rate-unrelated frequency ranges are not taken as heart rate and the heart rate is temporally dependent on previous ones.

However, this system has limitations including error propagation due to temporal dependencies of the algorithm and a lack of robustness to changes in signal properties. It was also unable to reconstruct the clean audio, meaning that the data could not be used for metrics other than heart rate. Thus, we acknowledge that a more sophisticated approach to the problem, specifically to addressing signal denoising, is required.

3.3.2 Overview of the deep learning-based pipeline

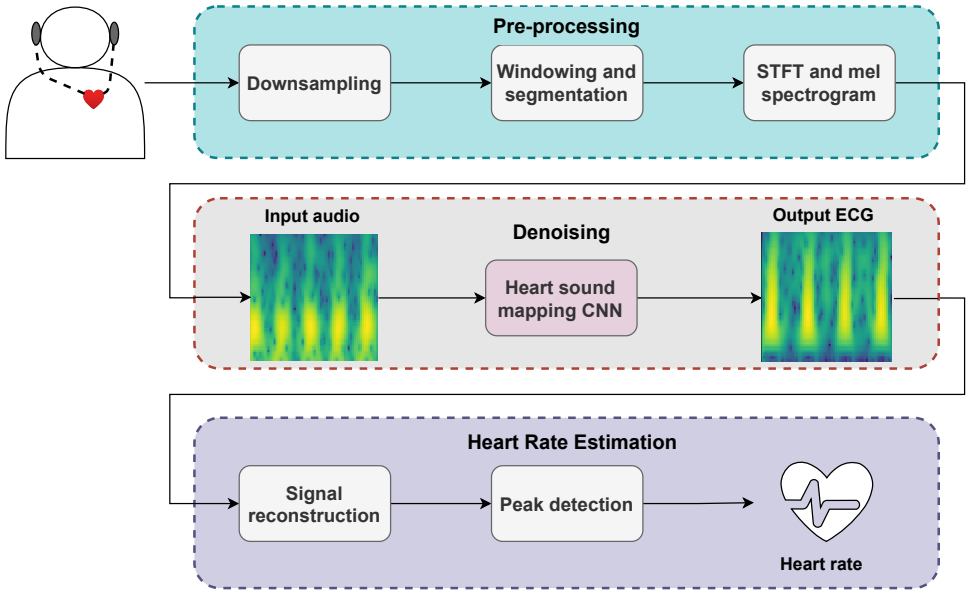


Figure 3.3.1: hEARt system flowchart.

An overview of hEARt, our designed motion-resilient heart rate monitoring system, is given in Figure 3.3.1. Audio signals captured inside the occluded ear canal are used for

heart rate estimation, which is performed in three stages: pre-processing, motion artefact elimination and heart rate estimation. Pre-processing aims at removing the frequency components unrelated to heart sounds. For motion artefact elimination, we proposed a CNN-based network to map spectrograms of noisy heart sound signals to spectrograms of the corresponding ECG (a clean heart signal) during the training phase, thus producing an output *synthesised ECG*. Our problem is thus framed as a denoising problem, but also as a synthesis problem. We adopted a U-Net encoder-decoder architecture (as detailed in Chapter 2) for denoising since audio (and specifically heart sounds) is commonly represented in image form as spectrograms [126, 145, 146]. Initially developed for biomedical image segmentation, U-Net shows great potential in image denoising and super resolution [126,147]. It captures important features in audio spectrograms via an encoder and reconstructs the corresponding clean heart signal via salient representations via a decoder. More importantly, the skip connections in U-Net allow the reuse of feature maps to enhance the learning of the original information, making it suitable for denoising. Evidence shows that U-Net performs well with limited training data, which matches our case [126, 148], rather than complex network structures [149, 150] which easily suffer from overfitting. Finally, heart rate is estimated using peak detection on the clean signals.

3.3.3 Pre-processing

The heart sounds captured by the in-ear microphone are low-frequency signals with a bandwidth of less than 50 Hz. To prepare the audio signals for processing, we downsample the audio from 22 kHz to 1 kHz to reduce computational complexity and remove unnecessary frequency content. We then segment the audio into 2-second windows, each with a 1.5-second overlap with the previous window. 2-second windows were selected to ensure the presence of multiple heartbeats (at least 2) within a window, enabling the system to learn inter-beat properties. Each window is bandpass filtered between 0.5 Hz and 50 Hz using a fourth-order Butterworth filter to remove the DC offset and high-frequency signals. This attenuates the frequency components not of interest for heart rate calculation, including music and ambient noise. Additionally, due to the occlusion of the ear canal, the majority of external noise is suppressed and not captured by the internally facing microphone. However, as outlined in Section 3.2.2, motion artefacts and other interfering signals lie in overlapping frequency ranges with heart sounds, therefore requiring additional processing.

We process the ground truth ECG similarly. The ECG, sampled at 250 Hz, is bandpass

filtered between 10 and 50 Hz and upsampled to 1 kHz. The highpass cutoff for the ECG was selected to be 10 Hz as this was empirically found to emphasise the peaks in the ECG (the QRS complex) while attenuating the P and T waves (as seen in Figure 3.2.1). Since we are only interested in capturing the beats and the inter-beat timing (for measuring heart rate, and in future, heart rate variability), only the QRS complex is of interest.

3.3.4 Motion artefact elimination

The motion artefact elimination subsystem takes as input pre-processed audio signals, and produces cleaned heart signals. To do so, it uses the ground truth ECG signal to supervise the denoising of the heart signals.

Spectrogram generation

We compute log-mel spectrograms (a representation of how a signal changes in frequency over time based on the mel scale, which mimics the way humans perceive sound) of the windowed audio and ECG signals using the STFT, with a window size of 256 samples and a hop length of 32 samples. 1024 FFT bins are used with zero padding and a Hann window. These parameters were empirically selected to clearly visualise heart sounds in the resulting spectrogram. Thereafter, the log-mel spectrogram is computed using 64 mel bins. Log-mel spectrograms were chosen over spectrograms since they provide more detailed information in the low-frequency region, where heart sound frequencies reside. The resulting log-mel spectrogram is a 64×64 matrix for each window. Since audio is captured in both ears, a spectrogram is computed for each channel and stacked together to form one 64×64×2 input. The output is a single-channel ECG spectrogram. The spectrograms are normalised between 0 to 1 using min-max normalisation to aid network training.

Deep learning architecture

Figure 3.3.2 provides the architecture of the denoising U-Net. In the encoder (or contraction path), the model consists of repeated 3×3 convolutions (with a ReLU activation function), batch normalisation and max pooling blocks with a stride of 2 to downsample the data. After pooling, dropout is applied with a rate of 0.1 to avoid overfitting. Each time the data is downsampled, the number of feature maps is doubled to enable the network to learn complex structures in the data. In the decoder (expansion path), the data undergoes successive up-convolutions where the number of feature maps is halved at each step. After

each up-convolution, the feature maps are merged with the corresponding map from the encoder and then undergo convolution and batch normalisation layers as in the encoder. In the final layer, a 1×1 convolution maps the feature maps into a single 64×64 output image.

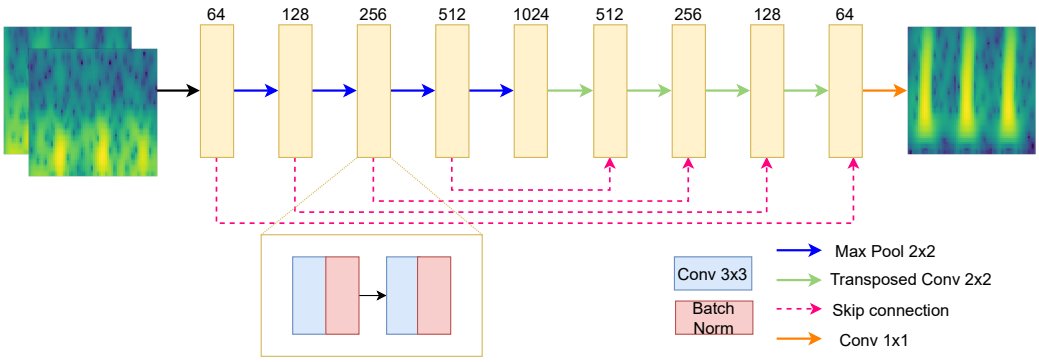


Figure 3.3.2: U-Net autoencoder architecture.

Transfer learning

On account of the small dataset, transfer learning is used to improve the results of the heart sound denoising. To achieve this, the model is pre-trained using the PASCAL heart sounds dataset [151], a dataset of heart sounds of both healthy participants and participants with murmurs which was collected using an iPhone and a digital stethoscope. When pre-training, we use log-mel spectrograms of heart sounds as both input and label to the network. By doing this, we aim to improve the ability of the network to extract representative audio features and encodings related to heart sounds. The pre-trained model weights are set as the initialisation weights for the CNN, which is fine-tuned using our data. This helps leverage and transfer the knowledge learnt about heart sounds using PASCAL, and also helps to avoid overfitting on a small dataset.

Training

The input audio spectrograms and their corresponding ECG spectrograms are used to train the network. We use leave-one-out cross-validation for testing whereby each subject is held out as the test set and a model is trained on the other 14 users. The model is trained for 100 epochs (empirically selected) using the Adam optimiser with a learning rate of 0.001 and batch size of 128. When choosing training parameters, our objective was to strike a good balance between performance and computational complexity.

hEARt uses the structural similarity index measure (SSIM) as the loss function [152]. SSIM is a measure of image similarity, which takes into account the local structure of the image. SSIM has been successfully used for image reconstruction and image denoising [152], making it suitable for our task. SSIM loss ($L_{\text{SSIM}}(P)$) is defined in Equation (3.1), where the SSIM (SSIM(p)) of two windows x and y is defined in Equation (3.2). In Equation (3.2), N is the number of pixels (p) in a patch (P), μ and σ are mean and standard deviation respectively, and C_1 and C_2 are constants [152].

$$L_{\text{SSIM}}(P) = \frac{1}{N} \sum_{p \in P} 1 - \text{SSIM}(p) \tag{3.1}$$

$$\text{SSIM}(p) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.2}$$

Signal reconstruction

We convert the reconstructed clean spectrograms to time-domain waveforms for heart rate estimation. The Griffin-Lim algorithm [153] is used for spectrum inversion due to its ability to reconstruct signals from spectrograms without phase information. The converted waveforms are then merged into a continuous time-series signal by averaging the overlapping regions.

3.3.5 Heart rate estimation

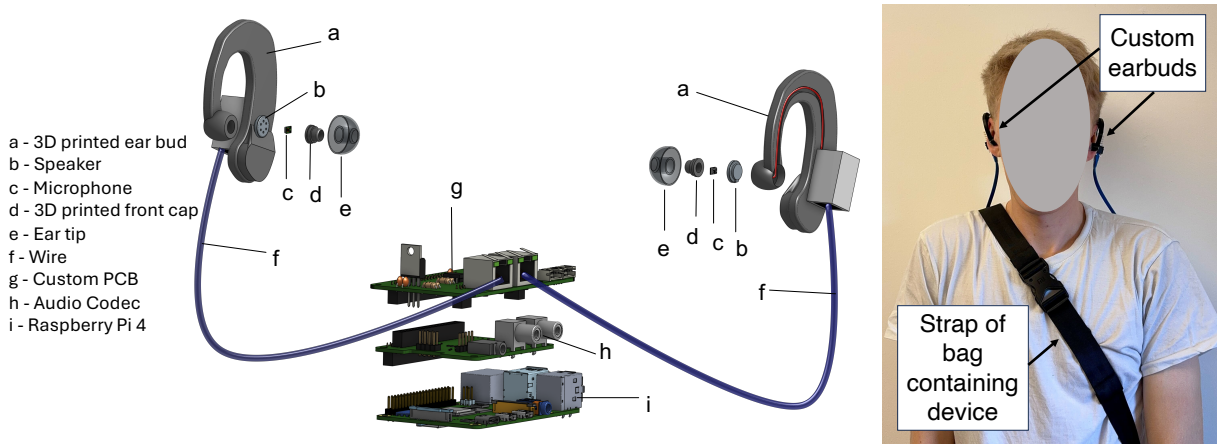
Heart rate estimation is performed on a 10-second long window, where each window has a 6-second overlap with the previous window [154]. Each window undergoes the Hilbert transform to compute the envelope of the signal. Thereafter, a Gaussian moving average filter smooths out small ripples and peaks in the signal. Peak detection is performed on the resultant signal, and the timings between consecutive peaks are used to compute the average heart rate for the window. Finally, a moving average window of 5 samples is used to remove outliers from the predictions. 5 samples were used to obtain accurate results, while still ensuring that deviations and changes in heart rate were evident.

3.4 Implementation

In this section we present the implementation details of our system, describing our prototype and the methodology we followed to run our data collection campaign.

3.4.1 Prototyping

While in-ear microphones have been successfully integrated into existing commercial earbuds such as the Apple AirPods Pro [6], no API currently exists to access the in-ear microphone output from these devices. To gather data and gain insights into the potential of our approach, we developed a custom earbud prototype (Figure 3.4.1(a)). We 3D printed an earbud in an ear-hook shape to ensure a secure attachment to the ear while also allowing space to mount additional sensors and the necessary electronic components. Inside the eartip, we embedded a Knowles SPU1410LR5H-QB microphone [155] selected for its flat frequency response between 10 Hz and 10 kHz, thus enabling the capture of low-frequency bone-conducted sounds and heart sounds. The microphone was secured inside the earbud facing towards the ear canal. We also embedded a speaker behind the microphone to enable audio playback.



(a) (b)

Figure 3.4.1: Custom earable device. (a) Exploded schematic of the custom earable system for in-ear audio collection and (b) participant wearing the device.

To record signals from the microphones, we designed a PCB which interfaces with an audio codec [156] onto a Raspberry Pi 4B. The PCB contains an MCP6004 non-inverting operational amplifier with adjustable gain which is controlled with potentiometers, allowing

the signals to be amplified before being sampled by the audio codec onto the Raspberry Pi. Further detail on the hardware design is provided in Appendix A.

To make the system portable, we placed the Raspberry Pi and PCB into a chest-worn bag and powered it with a portable power bank, as shown in Figure 3.4.1(b). This ensures that the prototype did not impede the natural movement of the participant. During walking and running, we secured the chest bag to the participant using a velcro strap to prevent it from bouncing while undergoing motion, thus limiting motion artefacts in the signal caused by the movement of the wires.

3.4.2 Data collection

We used a Zephyr BioHarness 3.0 chest strap [65] to measure the ground truth heart signal. Specifically, we extracted the raw ECG from the Zephyr BioHarness and used it as both the clean heart signal for the CNN and to calculate the ground truth heart rate. The microphone data was sampled at 22050 Hz and the ECG at 250 Hz. As a result of the different sampling rates, there is a maximum of a 90 ms delay between the audio and the ECG signal. However, since heart rate estimation is performed in 10-second windows, this delay is negligible.

We invited 15 participants (8 males and 7 females) for data collection which was approved by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge. The average age of the participants was 31±8, with ages ranging from 23 to 55. We ran an extensive data collection protocol consisting of sedentary and motion scenarios, as shown in Table 3.4.1, amounting to a total of 41 minutes of data per participant. These activities were selected as they encompass typical scenarios when a person uses earbuds and requires heart rate monitoring. Our data collection involved controlled activities and also in-the-wild scenarios to ensure the applicability of our methods to real-world use. For the controlled activities, data collection took place inside a laboratory environment containing a treadmill and a desk. Experiments were also conducted outdoors and in an office. We had to remove the data for participant 12 while walking due to poor ground truth data quality.

Motion activities were performed on a treadmill to control cadence, however, participants were given a choice of three speeds so that a comfortable pace could be selected. For walking, speeds ranged between 2km/h and 5km/h. For running, they ranged between 5km/h and 8km/h. The average sound level in the room was 32 dB while stationary and

Table 3.4.1: Data collection protocol.

Activity	Duration (minutes)	Mean heart rate	Mininum heart rate	Maximum heart rate
Sedentary				
Sitting	5	67	51	87
Standing	5	76	57	92
Lying down	5	61	45	84
Listening to music	3	79	56	97
Meditation	3	65	45	86
Working in the wild	5	66	48	87
Cool down after exercise	5	87	55	144
Moving				
Walking	5	91	72	108
Running	5	141	74	182

58 dB under motion.

The cool down activity was performed immediately after the user had completed their walking and running, and so captured their natural recovery heart rate. While sitting, users were asked to move their head three times to test the accuracy of heart rate estimation in the presence of non-full body motions. While meditating, participants were given a video to follow with controlled inhalation and exhalation durations to try to slow the heart rate.

Table 3.4.1 also provides the mean, minimum and maximum heart rate for each activity. The distribution of heart rates varies per activity, with running having the largest range of heart rates. These activities also clearly encompass a large range of heart rates, with heart rates of up to 144 BPM even while stationary. This dataset is thus well suited to test the ability of the system to predict heart rate under a wide range of conditions.

3.5 Evaluation

3.5.1 Metrics

We evaluated the performance of our system according to the following metrics [51]:

(i) Mean Absolute Error (**MAE**): the average absolute error between the ground truth heart rate (HR_{true}) and the calculated heart rate (HR_{calc}) for each window ($i, i \in [1, N]$)

(Equation (3.3)).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|HR_{calc}(i) - HR_{true}(i)| \tag{3.3}$$

(ii) Mean Average Percentage Error (**MAPE**): the average percentage error over each window (Equation (3.4)).

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\text{HR}_{\text{calc}}(i) - \text{HR}_{\text{true}}(i)|}{\text{HR}_{\text{true}}(i)} * 100 \tag{3.4}$$

(iii) Modified Bland-Altman plots: a scatter plot indicating the difference between the two measurements (i.e. the *bias* or error) for every true value (i.e. heart rate from the ground truth). A modified Bland-Altman (BA) plot is constructed so that 95% of the data points lie within ±1.96 standard deviations of the mean difference between the methods [157]. BA plots are used clinically to assess the level of agreement between two measurement methods [157]. For evaluating performance, we compare the calculated heart rate with the ground truth heart rate for each 10-second window.

3.5.2 Baseline comparison

Table 3.5.1 shows the performance comparison between the proposed deep learning-based hEARt system and two signal processing approaches:

1. The signal processing method proposed in Section 3.3.1 (referred to as signal processing) which leverages the DWT for signal denoising and extracts heart rate from the frequency spectrum of the denoised signals.

2. The baseline developed by Martin and Voix [10] (referred to as baseline), which uses Hilbert transforms and peak detection for heart rate estimation in the time domain under stationary conditions. This baseline was selected as it is the most relevant to the work and commonly used baseline in the literature.

To perform this comparison, we group our activities into three scenarios based on the level of full body activity: **(1)** sedentary (comprising of sitting, standing, lying down, listening to music, meditation, working and cooling down after exercise); **(2)** walking; **(3)** running. Our proposed signal-processing approach outperforms the baseline for all activities. This demonstrates that the baseline algorithm designed for stationary is unable to generalise

to motion conditions, and an additional denoising module is required. Additionally, it shows that the motion artefacts introduced by in-the-wild stationary activities greatly deteriorate the performance of heart rate estimation, meaning that even while sedentary more sophisticated techniques are needed. Comparing the signal processing approach with hEARt, we observe that hEARt significantly outperforms signal processing for all activities, showing that the deep learning-based technique is better than the signal processing approach at generalising to differences in the data and motion artefacts.

Table 3.5.1: Comparison between hEARt, the two baselines and in-ear PPG in terms of MAPE (%).

Activity	**hEARt**	**Signal Processing**	**Baseline [10]**	**In-ear PPG [58]**
Sedentary	**2.82±4.81**	10.99±14.52	18.01±12.65	—
Walking	**7.78±6.17**	19.31±13.41	22.74±17.39	27.14
Running	**9.60±9.28**	20.39±14.07	25.67±11.09	29.84

With more intense, full-body motion interfering with the heart sounds (as with walking and running), the signal processing approach fails to accurately capture the heart rate from the signal and the performance severely deteriorates. hEARt outperforms signal processing significantly with a relative improvement of 60% and 53% for walking and running respectively, suggesting the effectiveness of hEARt in heart rate estimation.

We also compare the performance of hEARt with that of in-ear PPG (as studied by Ferlini et al [58]) in Table 3.5.1. It is evident from the table that (i) although PPG is the gold standard for heart rate measurement, full-body motion causes significant degradation in heart rate measurement quality and (ii) our audio-based approach performs better than in-ear PPG. We thus believe that in-ear audio could be used as an alternative to, or in combination with, in-ear PPG for heart rate measurement through the ear.

3.5.3 hEARt overall performance

The overall performance of hEARt in terms of both MAE and MAPE for each activity are provided in Table 3.5.2. The average MAE of the sedentary activities is 1.88±2.89 BPM. From the table, it is evident that, as expected, full-body motion degrades heart rate estimation ability with errors increasing with increasing intensity of motion. By examining the sedentary activities, we also see the impact of motion artefacts on estimation error. Specifically, working in the wild has the largest error of the sedentary activities, which

Table 3.5.2: Performance of hEARt for each activity in terms of MAE (BPM) and MAPE (%).

Activity	MAE (BPM)	MAPE (%)
Sitting	1.37±1.59	2.11±2.61
Standing	1.50±1.78	2.01±2.50
Lying down	2.00±3.6	3.68±7.22
Listening to music	1.67±2.04	2.25±3.01
Meditation	2.33±2.71	3.95±4.94
Working in the wild	2.47±3.83	4.07±6.68
Cool down after exercise	1.96±2.98	2.24±3.06
Walking	6.83±5.05	7.78±6.17
Running	13.19±11.37	9.60±9.28

is intuitive since any in-the-wild scenario will inherently contain more movement (thus more artefacts in the signal) than a controlled scenario. The error increases for walking and running with MAE of 6.83 BPM and 13.19 BPM respectively. However, errors for all activities are less than 10%, meaning that our system is accurate by ANSI standards [138].

One notable result is the excellent performance while playing music through the speaker on the earable device. Even in the presence of music, heart rate estimation is accurate with a MAE of 1.67 BPM. This proves that hEARt can be used even during music playback, thus fulfilling one of the key use cases of an earable device.

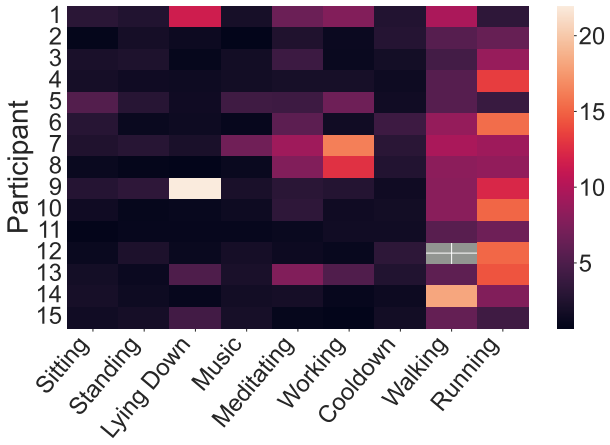
3.5.4 Individual heart rate estimation

Next, we evaluate our approach under different activities for all participants. First, we provide some insights into the population statistics. Figure 3.5.1(a) reports a heatmap of the MAPE of the audio-extracted heart rate for every user across the activities. Lighter colors correspond to larger MAPE values. Walking for user 12 was removed due to poor quality ground truth ECG, and is represented by a grey box (or NaN error). From the figure, we can extract a number of insights:

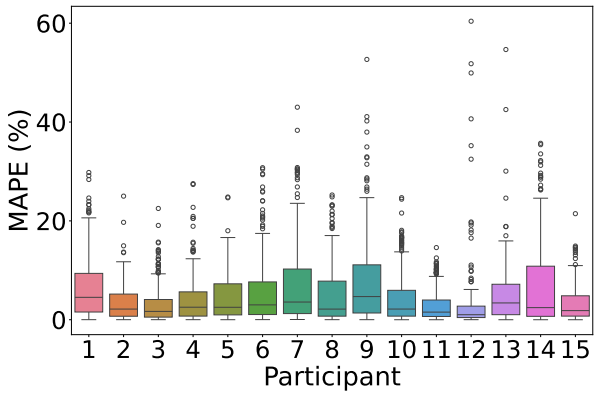
- Errors for motion conditions are higher than stationary.
- Our system generalises well to the different activities with mostly consistent errors amongst participants for the activities.
- Certain users experience poor performance in a specific activity (e.g. user 9 for lying

down and user 14 for walking). This is likely due to an incorrectly fitting earbud in one ear which loosened during the activity, reducing the occlusion effect. These issues would be solved by the use of wireless earbuds (ensuring that the wires do not dislodge the earbuds during activity) and by ensuring good fit quality of the earbuds for each user.

Overall, these results prove that the system is able to generalise to different users and that with high-quality data, good heart rate estimation can be achieved.



(a)



(b)

Figure 3.5.1: (a) Heatmap of the mean absolute percentage error per participant and (b) boxplot of mean absolute percentage error per participant.

To further assess the performance of hEARt on individuals, we provide a boxplot of the MAPE of heart rate estimation across all activities for each participant in Figure 3.5.1(b). It is evident that there is little variation in the median error of each participant, with the median error ranging between 1% for participant 12 and 4.7% for participant 9. The

two participants with the largest median error (as shown by the vertical line through each box) are participants 1 and 9, while the participants with the biggest spread of errors are participants 9 and 14. If we compare the heatmap and the boxplot, we can see that participant 1 has larger errors for lying down, meditating, working and walking resulting in a higher median error. Participant 9 has a very large error for lying down which results in a larger median error and a significant number of outliers. This error is due to weak signals when lying due to the displacement of the earbuds. We can also see that participant 12 has a small median error but a large number of outliers. This is on account of the large error in the running activity, where a few windows were incorrectly estimated with large errors. However, overall, the results show that the system is capable of generalising across different users and activities.

To further understand the extent to which the various activities impact hEARt, we report the cumulative distribution function (CDF) of the error for each activity in Figure 3.5.2. These results align with those in Figure 3.5.1, indicating that our approach achieves under 10% error for over 60% of windows across all users and activities. Most errors stem from specific users rather than the general population, as highlighted in the boxplot. This performance on our academic prototype confirms that in-ear audio-based heart rate monitoring offers a promising modality for continuous heart rate sensing in the presence of motion.

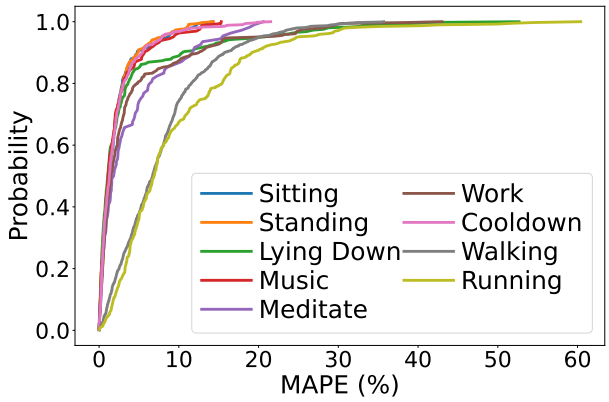
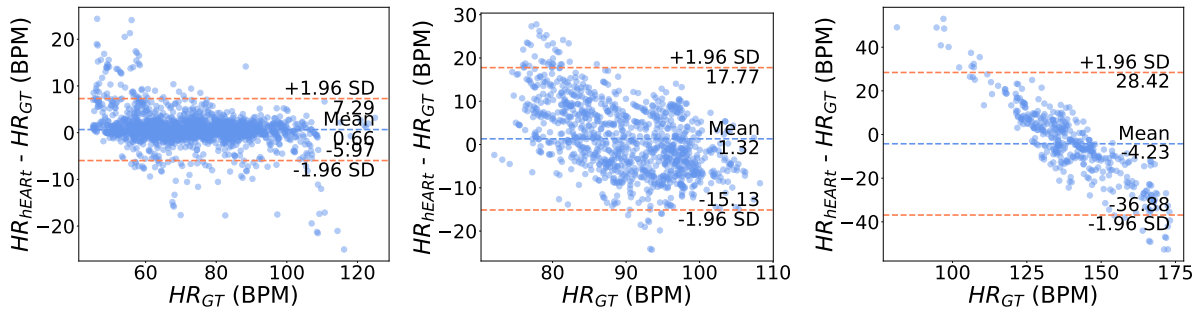


Figure 3.5.2: CDF of the mean absolute percentage error of heart rate estimation over the different activities.

3.5.5 Bland-Altman plots

To further analyse the results, we leverage modified BA plots. We report BA plots (i.e. the agreement between the heart rate calculated with hEARt and that obtained from the GT



(a) Sedentary (b) Walking (c) Running

Figure 3.5.3: Modified Bland-Altman plot of heart rate (HR) extraction.

chest strap) for the three scenarios based on activity level (stationary, walking, running) in Figure 3.5.3. Specifically, Figure 3.5.3(a) reports the agreement while stationary. It is clear that the bias between the two measurements is minimal, with very low mean error (only 0.66 BPM) and narrow limits of agreement (dashed red lines). Notably, the majority of the data points fall inside the limits of agreement, denoting the two measurements are in agreement. On the other hand, with more intense activities like walking and running (Figure 3.5.3(b) and Figure 3.5.3(c) respectively), wider limits of agreement are present, representing a greater standard deviation in the error in heart rate estimation. For walking, the positive mean error shows a slight overall tendency for overestimation. Additionally, there is a slight trend where lower heart rates are overestimated and higher ones are underestimated, however, the errors are quite centrally distributed. However, for running this trend is very distinctive, where there are larger errors at higher heart rates and a tendency to slightly underestimate heart rate. This can be traced down to the imbalance of our dataset, where lower heart rate values are predominant, due to the larger number of sedentary tasks. Especially in the running case, this underestimation of higher heart rates is observed when the frequency of the running overlaps with the heart rate values. The motion artefact-induced spikes trigger a harsher response by hEARt which tries to remove the noisy peaks, thus leading to an underestimation of higher heart rates. Nonetheless, our approach still performs well for all activities.

3.5.6 Heart rate estimation while speaking

One key use case for an earable heart rate monitoring system is for estimating heart rate while speaking. To assess this, we collected data from a subset of 13 participants while speaking out loud for 2 minutes. Under this setting, we achieve a MAE of 7.51±5.29 BPM,

amounting to a percentage error of 9.97±6.42%. Thus we see that the error for speaking is still less than 10%, however, there is more variation in this error than for the other activities listed in Table 3.5.2. Perhaps against intuition, this higher error is not on account of speech being detected by the microphone since the frequencies of audible human speech are significantly higher than those of interest in the hEARt system. Rather, speaking causes movement of the jaw and head, and deformation of the ear canal due to jaw movement. These movements result in low-frequency bone-conducted vibrations which could be interpreted as heartbeats. They are also non-periodic and random in nature and thus harder to remove, resulting in higher errors. This is in contrast to walking and running which are largely periodic and more homogeneous and thus easier to remove. We also have limited data collected for the speaking activity, meaning that the model has limited training data from which to learn the random patterns of speech. It is expected that with more data collected for longer durations and from more participants, this error will decrease.

3.5.7 Outdoor performance

To ensure the applicability of our system to real-world scenarios and diverse ambient noise levels, we assessed performance for three users outdoors. During these tests, we specifically evaluated four activities typically done outdoors: sitting still, walking, running, and cooling down.

The tests took place near an active building construction site, where the ambient sound level averaged 53 dB. During the walking and running assessments, participants were allowed to select their preferred pace and move freely within the designated area. This setup was designed to replicate real-life conditions and variations in noise levels, making our evaluation more comprehensive and reliable.

Table 3.5.3 provides the results for the outdoor study. By comparing with Table 3.5.1, it is evident that the results obtained outdoors are consistent with those obtained in the controlled laboratory setting, showing the robustness of the proposed system. This result shows that not only is hEARt able to accurately predict heart rate in the presence of external noise, but also validates that the system works while walking and running both on the treadmill and freely on the ground.

Table 3.5.3: Performance of hEARt in outdoor settings.

Activity	MAPE (%)	SAPE (%)
Sedentary	**1.90**	**3.64**
Sit	1.84	4.36
Cool down	1.96	2.91
Motion	**11.07**	**7.76**
Walk	11.77	7.86
Run	10.37	7.66
Average	**6.49**	**5.65**

3.5.8 Long-term tracking performance

The results of the previous sections were obtained from experiments conducted under controlled conditions. To assess the real-world effectiveness of the designed system, we collected an hour of data from three participants under conditions of daily life. The tests were done in a busy shared office with an average sound level of 43 dB. During this time, the participants were instructed to undergo their activity as normal. At the time, participants were at work in an office, and so this activity included walking around the office and sitting and standing at their desks while working thus resulting in the participants performing uncontrolled movements. The longitudinal tracking results of heart rate prediction for one user in this study are given in Figure 3.5.4. From the figure, it is clear that the system is able to accurately predict heart rate even in uncontrolled environments as the trends of the two lines very closely match one another. We see slight underestimations of the heart rate while walking, however, the overall estimations are accurate.

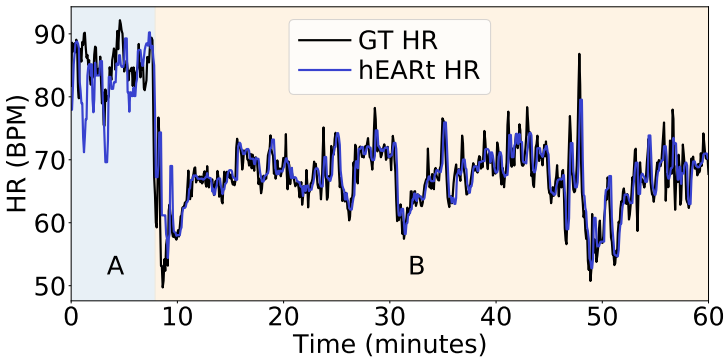


Figure 3.5.4: Longitudinal heart rate tracking using hEARt. Coloured boxes indicate the different activities. A: Walking. B: Working in the wild.

The MAE of this longitudinal study is 2.83 BPM, corresponding to a MAPE of 4.07%. To analyse this further, the MAPE of activities A and B (i.e. walking and working in the wild) are 7.26% and 3.79% respectively. If we compare these results to those in Table 3.5.1, we can see that the activities have comparable performance to that of the controlled experiments. The results of this study prove that the model is generalisable to different conditions and activities. Thus this study acts as a proof of concept of the in-the-wild feasibility of hEARt.

3.5.9 Power and latency measurements

For a comprehensive system analysis, we evaluated the power consumption and latency of our system when implemented on a Raspberry Pi 4. The trained hEARt CNN was converted to TensorFlow lite and deployed on the device. This mimics a stand-alone earable system whereby processing is done on-device. Table 3.5.4 provides a breakdown of the operation times for the various system components. Signal denoising was performed on a 2-second window and heart rate extraction on a 10-second window, as detailed in Section 3.3. Processing a 10-second window takes the system 347 ms, implying that a new heart rate can be predicted by the system every 4 seconds (due to the 6-second overlap). This latency parameter is adjustable based on the desired overlap ratio.

Table 3.5.4: Latency of hEARt for heart rate measurement.

Operation (window)	**Latency (ms)**
Preprocessing (2s)	1.66
Denoising (2s)	7.66
Reconstruction (10s)	17.96
Heart rate extraction (10s)	55.78
Total (10s)	83.06

The system power consumption is given in Table 3.5.5. Overall, the system (including microphone sampling, denoising and heart rate prediction) consumes 491 mW. The microphone sampling runs continuously, but the hEARt system is only active for 83.06 ms for each estimate, and an estimate is made every 4 s. Thus, the average energy consumed per second is $(3032 - 2870)\,mW \times 1s + (3361 - 3032)\,mW \times 83.06\,ms/4 = 699\,mJ$. Although this energy consumption might appear substantial, it's worth noting that our system was implemented on a power-intensive Raspberry Pi without optimising for energy efficiency. By implementing the model on a low-power microcontroller, power consumption will be reduced. Additionally, when converting the denoising CNN to Tensorflow Lite, optimisations and

quantisation were not applied. The model can thus be further optimised to reduce energy consumption and latency, thus lowering energy expenditure. These optimisations will be required before hEARt can feasibly be used on a commercial earbud. Therefore, through these experiments, we have shown the lower bound of system performance. We show that even in this highly unoptimised case, acceptable power consumption and latency can be achieved.

Table 3.5.5: Power consumption of hEARt for heart rate measurement.

Operation	**Power (W)**
Raspberry Pi (Baseline)	2.870
Raspberry Pi+Microphone	3.032
Full system	3.361

3.6 Conclusion

To address the limitations and high estimation errors of heart rate under motion currently seen in wearable devices, we introduced *hEARt*, a pipeline for deep learning-based motion artefact removal and heart rate estimation from audio signals collected in the ear canal. Our approach presents, for the first time, heart rate monitoring under motion using in-ear audio. We achieve results that outperform baseline comparisons, and that are accurate according to ANSI standards for a heart rate monitor.

This chapter therefore demonstrated the potential of using in-ear audio together with novel processing techniques to monitor heart rate even under the challenging scenario of full-body motion. It also demonstrated the ability of the in-ear microphone to detect body-related sounds, such as the sounds of footsteps and heart sounds, through bone conduction. Building on these findings, in the next chapter, we explore another critical vital sign, respiratory rate, and investigate the feasibility of respiratory rate monitoring using in-ear audio even in the presence of inaudible breathing sounds.

Chapter 4

Robust Respiratory Rate Monitoring

4.1 Introduction

In Chapter 3, we explored, for the first time, heart rate monitoring using in-ear audio under motion conditions. In this chapter, we build on the insights from the previous chapter and extend our work to another critical vital sign: respiratory rate. Motivated by the research gap described in Section 2.4, we explore robust respiratory rate monitoring under a variety of real-life conditions and activities, even under inaudible breathing sounds.

As discussed in Chapter 2, respiratory rate is a fundamental vital sign that relays pivotal information about a person's health and fitness. Continuously monitoring respiratory rate can help detect respiratory issues such as asthma, or chronic obstructive pulmonary disease, and deviations from normal breathing patterns may indicate underlying health problems. Respiratory rate can also be indicative of mental health status and can provide valuable insights into stress and relaxation levels. During exercise, respiratory rate helps estimate aerobic fitness, and effort levels, and assess breathing efficiency, allowing athletes to refine their breathing techniques to maximise performance. During the cool down after exercise, respiratory rate provides insights into how quickly the body recovers from exertion which is indicative of cardiovascular fitness. Therefore, there is significant value in continuous monitoring of respiratory rate across various sedentary and active daily settings.

In Chapter 2, we showed that existing respiratory rate monitoring solutions designed to facilitate respiratory rate monitoring in daily life primarily rely on three principles:

1. Monitoring breathing-induced body movements. Methods that use IMUs in smartwatches [71, 72, 74], smartphones [68–70], and earbuds [92–94] require the user to remain still. Pressure sensors in chest straps [65] necessitate the use of additional obtrusive wearable devices. Some systems use cameras [67] or acoustic sensing [66] on smartphones, requiring the phone to be held in a specific posture and the user to remain still.

2. Monitoring breathing-induced airflow through the nose or mouth. These methods involve obtrusive nose-worn sensors [158], microphones on smartphones positioned near the suprasternal notch and nose [11] or microphones on earbuds that detect only audible breathing sounds [10, 95, 159].

3. Using behavioural or physiological couplings. Previous studies [97, 98] exploit the relationship between respiratory rate and physical behaviours to indirectly estimate respiratory rate, but this method is limited to specific running conditions. Other studies [84–87] have estimated respiratory rate from physiological signs, but these only work at rest (such as while sleeping).

While each existing respiratory rate monitoring solution performs well under specific conditions, none can effectively monitor respiratory rate across a wide range of activities. Moreover, existing solutions often rely on constraints that are not manageable in daily life, such as user stillness or specific device positioning, which makes them impractical for daily use. *Therefore, a new approach is needed – one that surpasses the limitations of existing technologies to provide continuous, non-obtrusive respiratory rate monitoring that works effectively across a range of daily activities.*

To address this, this chapter presents RespEar, an earable-based system for robust respiratory rate monitoring across both sedentary (*e.g.,* sitting, standing, working, cooling down after exercise), and active (*e.g.,* walking, running, rowing, and step aerobics) activities. Due to their pervasiveness in many daily life activities (*e.g.,* entertainment, exercise, and work) and proximity to respiratory organs, earables are a natural choice to achieve robust respiratory rate sensing. In addition, in-ear microphones offer unique opportunities to measure breathing-related signals (*e.g.,* breathing sounds, heartbeats, and footsteps as shown in Section 4.2.1), which, together with unique algorithms, enables our solution.

While designing RespEar, we faced a number of challenges:

1) Almost imperceptible breathing sounds. The intensity of breathing sounds is

minimal when the user is sedentary and overwhelmed by other sounds, like footsteps when the user is active (Section 4.2.1) Thus, directly estimating respiratory rate using breathing sounds is unreliable (Section 4.4). To address this, we proposed a unified respiratory rate monitoring system by leveraging the unique properties of in-ear audio for the following purposes:

- *Respiratory Sinus Arrhythmia-based respiratory rate monitoring:* When clear heartbeat sounds can be captured using the in-ear microphone (predominantly when the user is sedentary as shown in Chapter 3), we can derive heart rate variability (HRV) from in-ear audio. Respiratory rate is then indirectly estimated using the Respiratory Sinus Arrhythmia (RSA) based physiological coupling between cardiovascular activity and respiration, *i.e.,* the association between respiratory rate and HRV.

- *Locomotor Respiratory Coupling-based respiratory rate monitoring:* However, when clear heartbeat sounds are not available (*e.g.,* in the presence of footstep sounds), RSA-based solutions are hindered by unreliable and inaccurate HRV estimation (validated in Section 4.4). Therefore, when rhythmic footsteps are present (*i.e.,* when the user is active), we rely on the in-ear microphone to capture low-frequency footstep sounds, which we use to derive the stride rhythm. Alongside faint high-frequency breathing sounds, respiratory rate is estimated by leveraging the Locomotor Respiratory Coupling (LRC) based physical coupling between gait and respiration, *i.e.,* the interaction of respiratory rate with stride rhythm.

2) Accurate and reliable estimation. Based on the above system, several technical challenges must be addressed to achieve accurate and reliable respiratory rate estimation:

- *RSA-based solutions under varying respiratory rates:* Practically, respiratory rate can change over time, introducing variability in the association between respiratory rate and HRV. This poses challenges in decoupling their relationship. To overcome this, we formulated an optimisation problem to dynamically extract breathing signals from HRV signals, thereby adapting to these variations to enhance performance.

- *Respiration-related features extraction for LRC-based respiratory rate monitoring:* Although the LRC shows the synchronisation between the stride rhythm and the respiratory rate, the LRC is both variable and unknown. This prevents direct respiratory rate estimation from stride frequency alone. It is thus necessary to extract respiration-related features from in-ear audio which, when combined with stride rhythm, can be used to esti-

mate respiratory rate accurately. However, while the user is undergoing active conditions, faint breathing sounds are heavily interfered with by other sounds, such as footsteps. To precisely extract the breathing-related features and facilitate respiratory rate estimation, we propose a scheme to estimate the probability of each audio frame containing breathing. We then apply Singular Spectrum Analysis (SSA) [160] to the generated probability curve to isolate components related to breathing.

- *Varying LRC ratios:* Respiratory rate is typically estimated for a window. However, within a window, LRC ratios may vary, *e.g.,* in walking or running by non-regular runners. To address this variability, we propose a method that aggregates breathing-related components from SSA by considering a range of possible LRC ratios, rather than choosing a fixed one.

We implemented RespEar using the prototype introduced in Chapter 3, and deployed the system on an iPhone 12 Pro. We evaluated RespEar across 8 different activities involving 18 subjects, achieving an overall MAE of 1.71 BPM and MAPE of 9.68%, with errors of 1.48 BPM (9.12%) and 2.28 BPM (11.04%) under sedentary and active conditions, respectively. We compared RespEar with recent earable-based solutions with IMUs [94], outer-ear microphones [74,97], and in-ear microphones [10], and found that RespEar outperformed all related works in both sedentary and active conditions while additionally being able to cater for both sets of conditions, unlike recent works. Additionally, we tested RespEar under a range of realistic conditions such as different noise levels, with music playback, in outdoor and indoor environments, in the wild, and at different moving speeds.

In summary, this chapter makes the following contributions:

- We explore respiratory rate estimation with common sensors on earables and identify unique properties of in-ear audio for respiratory rate estimation.
- We propose RespEar, the first earable-based system offering continuous and non-obtrusive respiratory rate monitoring across diverse daily activities. RespEar leverages solely the in-ear microphone, a sensor naturally present in many earables, together with intrinsic relationships of our cardiovascular, gait and respiratory systems to estimate respiratory rate.
- We implement RespEar and describe our extensive dataset and evaluation. Our results demonstrate that RespEar outperforms the state-of-art and is uniquely able to generalise beyond what other systems have been able to do in terms of activity

intensity while remaining robust under different environmental conditions.

The techniques and results in this chapter have been published in [18].

4.2 Primer

4.2.1 Preliminary investigation: sensors on earables

We first explore the feasibility of the three most common sensors on earables for respiratory rate monitoring, namely, *IMU*, used for motion detection and interaction [2], *out-ear microphone*, used for speech capture [2], and *in-ear microphone*, used for active noise cancellation [5]. We simultaneously collected signals from these sensors when a subject was breathing naturally under three conditions, as shown in Figure 4.2.1: sitting still, walking, and running on a treadmill. The Zephyr BioHarness 3.0 chest strap [65] was worn to collect reference signals. The collected signals are provided in Figure 4.2.1.

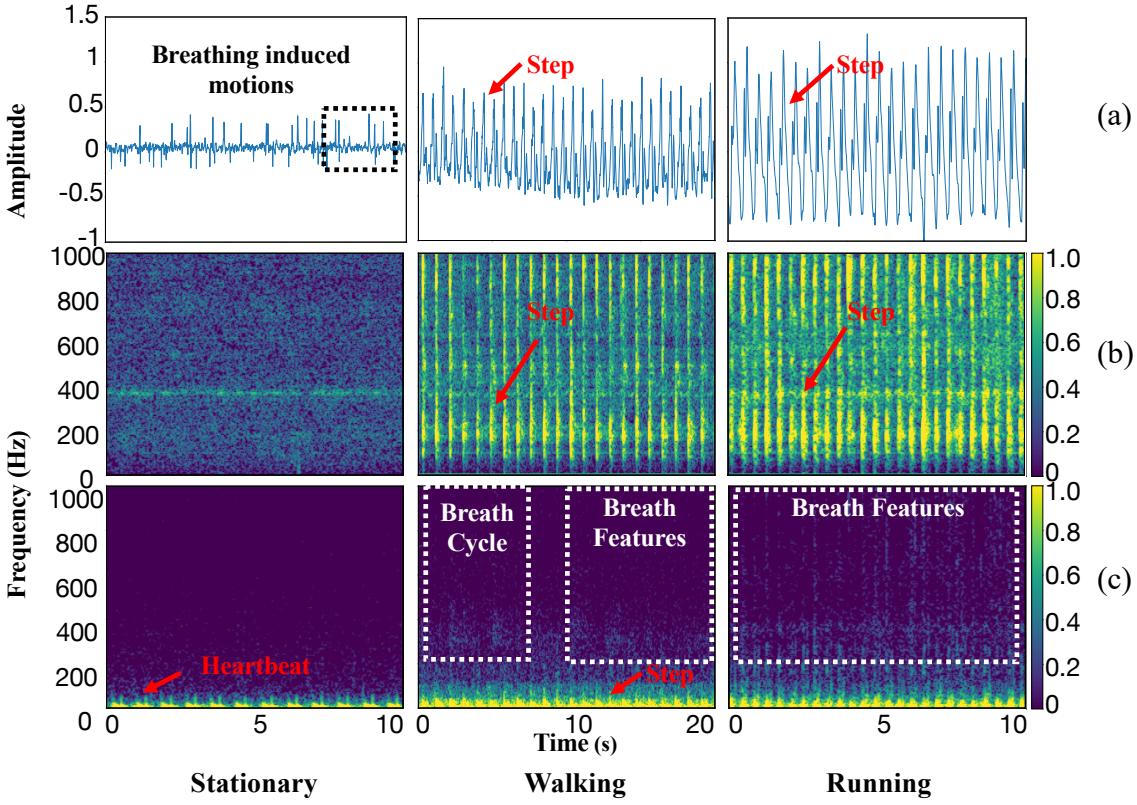


Figure 4.2.1: Signals captured using the (a) IMU, (b) out-ear and (c) in-ear microphone under different activities.

IMU

We observe from Figure 4.2.1(a) that IMUs can capture breathing-induced motions when the user is stationary, however, they are unreliable and have a low signal-to-noise ratio (SNR). Previous studies [92, 94] apply a bandpass filter on the IMU signals while stationary, achieving acceptable performance as validated in Section 4.4. However, in real-life, practical conditions, the subtle breathing signals are easily overwhelmed by head motions, resulting in poor data retention [93]. When walking or running, the IMU is clearly able to detect steps, but breathing signals are indiscernible since walking and running generate strong motions that completely overshadow the breathing signal [161].

Out-ear microphone

As per the spectrograms shown in Figure 4.2.1(b), the out-ear microphone captures only environmental noises under stationary conditions, failing to detect breathing sounds due to large attenuation of faint breathing in air. During walking or running, footstep sounds can be detected (depending on shoe type and ground material), however, breathing sounds are not discernible due to the attenuation of breathing in air, the loud footstep sounds, and the ambient noise of the treadmill.

In-ear microphone

In Figure 4.2.1(c), under stationary conditions, the faint breathing sounds are undetectable by in-ear microphones. However, sounds of heartbeats are clearly captured due to the occlusion effect [162] (as introduced in Chapter 2, and detailed in Chapter 3). When walking or running, we observe that 1) as seen in Chapter 3, footsteps can be clearly detected because of the occlusion effect [163]; 2) due to the variations in breathing intensity, the full breathing cycle often cannot clearly be distinguished, yet incomplete sounds of breathing (breathing features) are still present in the signal; 3) the in-ear microphone is resilient to ambient noise as it resides inside the ear canal.

Summary

The IMU works in stationary conditions but is sensitive to motion. The out-ear microphone is vulnerable to ambient noise under all conditions and merely captures noise while stationary. The capability of the in-ear microphone to capture heartbeats while stationary, footsteps and partly discernible breathing during walking and running shows the potential to offer

a solution for respiratory rate monitoring that works effectively in various conditions. Therefore, we select the in-ear microphone as the sensing modality in RespEar.

4.2.2 Respiratory rate and physiological couplings

In-ear microphones can capture versatile audio data in different conditions, yet the method to accurately correlate these signals with respiratory rate remains unclear. Inspired by the physiological couplings between cardiovascular activity and respiration [84], *i.e.,* RSA, we initially explore methods grounded on this physiological principle.

RSA-based respiratory rate estimation

RSA is the natural variations in heart rate that occur due to synchronisation with the respiratory cycle [84]. Due to RSA, heart rate increases during inhalation and decreases during exhalation, resulting in a breathing-related modulation of the HRV (*i.e.,* the variation in the time interval between successive heartbeats) [164]. Rhythms in the low frequency (LF) range of the HRV, spanning from 0.04 to 0.15 Hz, serve as indicators of sympathetic modulation [165]. Those within the high frequency (HF) range (0.15 to 0.4 Hz) encapsulate rhythms governed by parasympathetic activity, which is closely related to respiration [165].

The strong heartbeat sounds from in-ear microphones under sedentary conditions offer the possibility of monitoring respiratory rate using RSA. However, while RSA exists under intense full body motion [166] (*i.e.,* in active conditions), accurately extracting heartbeat locations from in-ear audio for HRV estimation in such conditions remains an unsolved and challenging issue [16, 46, 162]. This was shown in Chapter 3 where although accurate heart rate estimation can be achieved within a 10-second window, smoothing and outlier removal is necessary for accurate results. However, for HRV determination, accurate positioning of each heartbeat peak is required, which is a more challenging task. We further validate the difficulty in HRV estimation under motion for breathing rate extraction in Section 4.4.

LRC-based respiratory rate estimation

We observe that audio from in-ear microphones under active conditions, such as walking, running, or other activities with rhythmic footsteps, is dominated by footstep sounds and full breathing cycles are not always discernible (Section 4.2.1). Thus, we explore the physiological couplings between gait and respiration.

LRC is a universal phenomenon in activities that produce and utilise energy rhythmically [167], such as walking, running, swimming, and rowing [168–171]. It demonstrates the interconnected dynamics between respiratory rate and stride rhythm [172], indicating the synchronisation between an individual's stride rhythm and their respiratory rate. This implies that there will normally be a certain number of steps for each breath (*i.e.,* inhalation or exhalation). LRC has been shown to improve breathing efficiency and reduce energy expenditure when stable step-breath ratios are used during locomotion [173]. In human locomotion, a number of LRC ratios are observed, *e.g.,* 4:3, 3:2, 2:1, where an LRC of 2:1 means two steps are taken for one breath. Thus, the in-ear audio, containing clear footsteps and partly discernible breathing sounds, offers the possibility for estimating respiratory rate based on LRC.

4.3 System Design

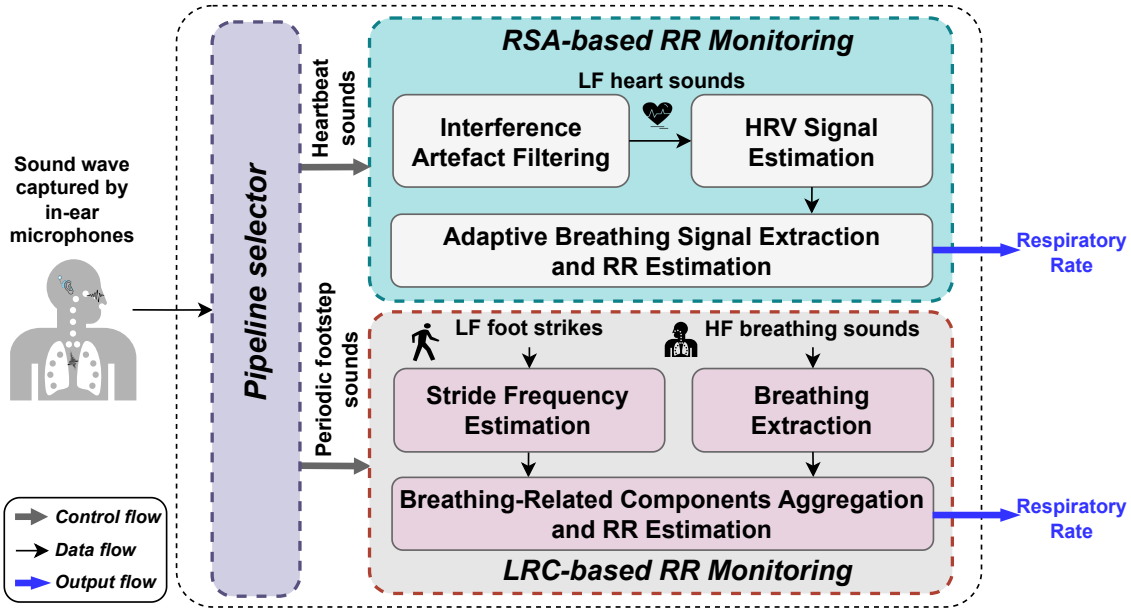


Figure 4.3.1: Illustration of the RespEar architecture.

Figure 4.3.1 illustrates the high-level architecture of RespEar. The system processes 60-second windows of in-ear audio, with a 30-second overlap between adjacent windows. Each window produces a single respiratory rate estimate. These window lengths were empirically

selected, but are adjustable parameters of the system. RespEar uses two paths for respiratory rate estimation, depending on the presence of rhythmic footsteps or clear heartbeat sounds in the audio. If the in-ear audio contains clear heartbeat sounds (*i.e.,* sedentary conditions), the respiratory rate is estimated through the **RSA-based respiratory rate monitoring** pipeline. If rhythmic footsteps are present in the in-ear audio (i.e., active conditions), respiratory rate is estimated using the **LRC-based respiratory rate monitoring** pipeline.

4.3.1 RSA-based respiratory rate monitoring

Design principle

The high-level process of RSA-based respiratory rate monitoring can be summarised in three steps:

1. **HRV signal estimation**: Heartbeats are detected, and the HRV signal is computed as the time difference between successive heartbeats.
2. **Breathing signal extraction**: A bandpass filter is applied to the HRV signal to capture respiration-related rhythms, extracting the high-frequency (HF) range as the extracted breathing signal.
3. **Respiratory rate estimation**: The final respiratory rate is determined by either applying peak detection to the extracted breathing signal and counting the peaks or by using the FFT (as detailed in Chapter 2) to identify the frequency component with the largest peak.

Prior RSA-based methods following the above process, primarily applied to PPG or ECG signals [84, 86–88], typically use default and fixed cutoff frequencies for the bandpass filter in step 2. However, we observe that using a fixed frequency range leads to sub-optimal or inaccurate localisation of the HF range in the HRV signal and therefore poor respiratory rate estimation. Since the HF range should be centred around the true respiratory rate, when the true respiratory rate changes over time, the HF range should change accordingly [165].

We conducted a study to validate this observation using all of our collected in-ear audio data with clear heartbeat sounds (*i.e.,* while sedentary). The distribution of the ground truth respiratory rate (denoted as $\mathrm{RR}_{\mathrm{GT}}$) for this data is shown in Figure 4.3.2(a). We follow the above process for RSA-based respiratory rate monitoring, whereby we determine respiratory rate by counting the peaks of the extracted breathing signal. The HRV signal is

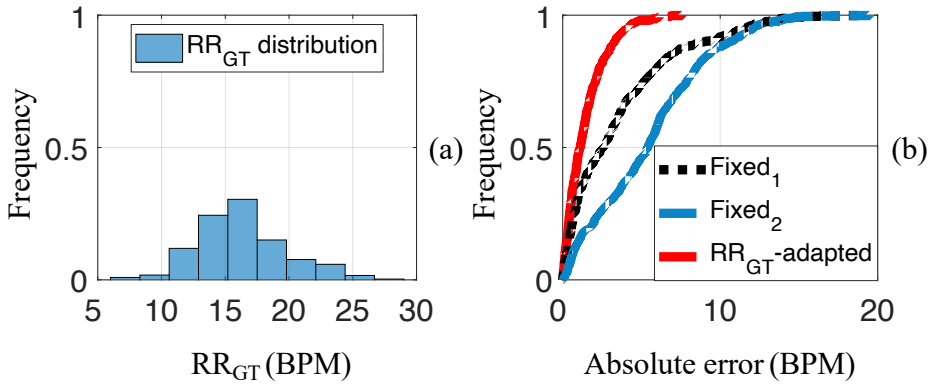


Figure 4.3.2: (a) Distribution of ground truth (GT) respiratory rate while sedentary. (b) Comparison of respiratory rate estimation performance using different bandpass filters.

calculated as the timing between heartbeats as detailed in the coming sections.

We compare the respiratory rate estimation performance using the following three filter ranges:

1. The bandpass filter using the default and fixed frequency range in $[87, 174]$, *i.e.*, $Fixed_1 = [0.15, 0.35]Hz$.
2. The predetermined bandpass filter covering the full frequency range of the RR_{GT}, *i.e.*, $Fixed_2 = [0.1, 0.5]Hz$.
3. The bandpass filter using an adaptive frequency range, *i.e.*, RR_{GT}-adapted, calculated as $[0.65 * \text{RR}_{\text{GT}}, 1.35 * \text{RR}_{\text{GT}}]$ [165].

The resultant performance is shown in Figure 4.3.2(b). It is evident that there is a large performance gain by using the RR_{GT}-adapted frequency range (MAE = 1.45BPM) compared to the $Fixed_1$ (MAE = 3.54BPM), and the $Fixed_2$ (5.45BPM). This is because the fixed bandpass filter is effective only if the ground truth respiratory rate falls within the range of the filter, but even then, using a HF range that is not centred around the true respiratory rate can degrade the performance.

Using adaptive HF range localisation on the HRV signal therefore has the potential to significantly improve the performance of traditional RSA-based respiratory rate estimation. We propose a novel approach whereby we formulate and solve an optimisation problem to dynamically localise the HF range. To the best of our knowledge, RespEar is the first work to achieve dynamic HF range localisation for RSA-based respiratory rate estimation.

We believe our methodology could also benefit other RSA-based solutions using various sensing modalities, *e.g.,* ECG and PPG. We also propose a series of techniques to enable the full pipeline for RSA-based respiratory rate monitoring using in-ear audio in the coming sections. Specifically, our RSA-based respiratory rate monitoring pipeline contains three components: (1) interference artefact filtering; (2) HRV signal estimation; (3) adaptive breathing signal extraction.

Interference artefact filtering

While the body is undergoing non-full body motions (*i.e.,* the user is sedentary), in-ear audio is prone to interference artefacts. Examples of these include head motions (drinking, speaking *etc.*), motions of the arms, movement of the trunk, and respiratory-related movements (swallowing, coughing, *etc.*). It must be noted that breathing adapts to speaking (inhalation at syntactic pauses and exhalation during speech [175]), which eliminates the need for respiratory rate estimation while speaking. To ensure accurate sedentary respiratory rate estimation even in the presence of interference artefacts, we present an interference filtering approach.

Interference artefact detection: Without interference artefacts, in-ear audio maintains a consistent waveform with clear heart sounds and thus stable signal statistics over time. Conversely, artefacts, such as one-time head motions, cause significant statistical variations. Hence, we propose a statistics-based approach to detect the presence of artefacts. For each 60-second window, we segment the audio signal into 3-second segments to capture short, one-time motions. We compute the standard deviation of each segment (to measure signal dispersion) and if it is larger than an empirical threshold, this segment is marked as an interfered segment.

Adaptive filter: If an interfered segment is detected, we use an adaptive filter implemented using the recursive least squares (RLS) algorithm [176] to remove the interference artefact. The RLS algorithm recursively finds filter coefficients that minimise the least squares cost function with a reference signal. For the segment containing interference, we select the nearest segment without interference as the reference signal. After filtering, the interference artefacts of the segment are mostly removed, enabling reliable respiratory rate estimation (as validated in Section 4.4). The cleaned signals are used as input to the HRV signal estimation module.

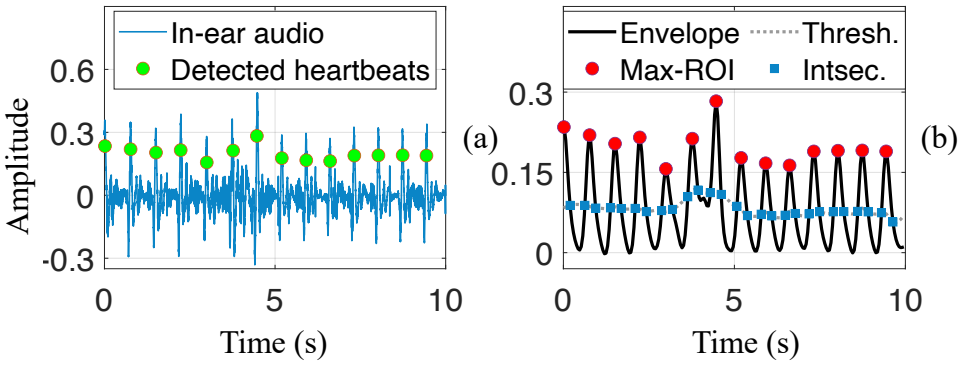


Figure 4.3.3: HRV signal estimation. (a) In-ear audio and heartbeats. (b) Peak detection using an adaptive threshold.

HRV signal estimation

Heartbeat detection: A low-pass filter with a 30Hz cutoff is applied to the in-ear audio to remove high-frequency noise and emphasise the heart signals. To derive the HRV signal (denoted as S_{HRV}), heartbeats are identified by detecting peaks in the filtered audio (Figure 4.3.3(a)). To accurately detect peaks while accommodating variations in amplitude and morphology changes in the in-ear audio, RespEar uses peak detection with an adaptive peak detection threshold.

As illustrated in Figure 4.3.3(b), the process begins by computing the Hilbert transform (described in Chapter 2) of the filtered audio. This is then processed with a moving average filter to generate the heartbeat envelope. Next, a moving average is computed for each point in the envelope, serving as an adaptive threshold (Thresh. in Figure 4.3.3(b)). Regions of interest (ROIs) are identified between the points where the envelope intersects the threshold. Heartbeat peaks are marked at the maximum point within each ROI (Max-ROI) between two intersection points (Intsec.), provided the amplitude is greater than that of the two surrounding intersection points. Finally, the HRV signal, S_{HRV}, is calculated as the time difference between successive detected peaks (Figure 4.3.4(a)).

Automatic channel selection: Leveraging the unique ability of earables to give two in-ear audio channels (*i.e.,* left and right ear), RespEar automatically selects and uses the channel with the lower standard deviation in the estimated HRV signal from each ear. This is because heartbeats are regular signals, so the lower standard deviation implies a less noisy and more robust signal due to more regular heartbeat peaks.

Adaptive breathing signal extraction

To adaptively localise the high-frequency (HF) range of the respiratory signal within the HRV signal (Figure 4.3.4(a)), we pose an optimisation problem based on two key observations:

1. Among a set of potential respiratory rate candidates, the best candidate is the one closest to the ground truth respiratory rate, $\mathrm{RR}_{\mathrm{GT}}$.
2. The optimal candidate should also minimise the difference between itself and the respiratory rate estimated by a bandpass filter centred around the candidate's frequency.

Thus, the objective is to select the respiratory rate candidate that minimises this difference, therefore providing the most accurate estimate of $\mathrm{RR}_{\mathrm{GT}}$. Consequently, we select the respiratory rate candidate with the minimum difference from the list of possible candidates.

The following steps describe the algorithm for adaptive breathing signal extraction, as detailed in Algorithm 1:

- **Identify the central respiratory rate candidate:** Perform a FFT on S_{HRV} to find the frequency component with the highest amplitude (Figure 4.3.4(a)).
- **Generate a list of respiratory rate candidates:** Sample potential respiratory rates around the central candidate to generate a list of candidate rates, $\mathrm{RR}_{\mathrm{list}}$.
- **Estimate respiratory rates using a bandpass filter:** For each candidate in $\mathrm{RR}_{\mathrm{list}}$, filter S_{HRV} using a bandpass filter with cutoffs defined by that candidate. Perform FFT on the filtered signal to estimate the respiratory rate by selecting the frequency component with the highest amplitude and converting the frequency to a respiratory rate (Figure 4.3.4(b)).
- **Calculate the frequency difference:** Compute the difference between each candidate and the respiratory rate estimated by the bandpass filter to create the frequency difference list (F-Difference-list) (Figure 4.3.4(c)).
- **Select the best candidates:** We observe that the optimal respiratory rate candidate typically appears among the three candidates that have the smallest local minima in the F-Difference-list. Choose these three candidates as the best respiratory rate estimates RR^{F}_{top} (Figure 4.3.5(a)).
- **Time-domain calibration:** To select the best respiratory rate estimate from the top

three candidates, we incorporate time-domain analysis. This improves the robustness of the estimation and resilience to noise and poor signal quality in S_{HRV}.

- Estimate the respiratory rate of each candidate in the time domain by counting the zero-crossing points in the corresponding filtered signal (Figure 4.3.4(b)).
- Compute the difference between the time-domain estimates and the candidate respiratory rates, generating a time difference list (T-Difference-list) (Figure 4.3.5(b)).
- To smooth out fluctuations and highlight underlying trends, we apply a smoothing filter to the T-Difference-list.
- Combine the F-Difference and T-Difference values for each of the three best candidates by summing the corresponding entries from both lists. The candidate with the smallest combined value is selected as the final estimated respiratory rate, $\text{RR}_{est}^{\text{TF}}$ (Figure 4.3.5(a,b)).

The output from this pipeline is the estimated respiratory rate per 60-second window while sedentary.

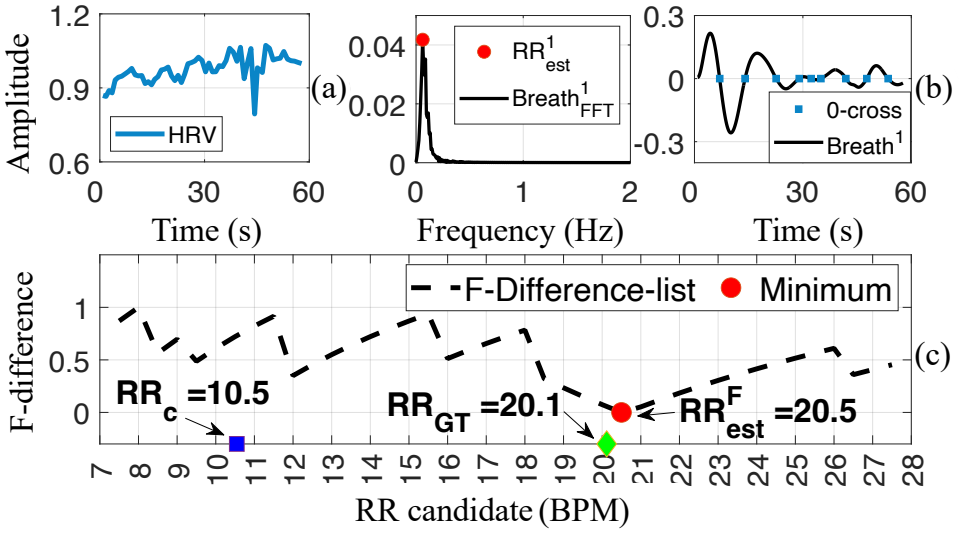


Figure 4.3.4: Adaptive breathing signal estimation. (a) Extracted HRV signal. (b) FFT computation and zero crossing counting. (c) Best respiratory rate candidate searching.

Algorithm 1: Adaptive Breathing Signal Extraction

Input: S_{HRV}: Heart rate variability signal
Output: RR_{est}^{TF}: Estimated respiratory rate

1. **Sample respiratory rate candidates (RR_{list})**
   1.1 **Calculating the centre RR of RR_{list}:** RR_c
   - Filter S_{HRV} using BPF with cutoffs $[0.15, 0.35]$ Hz
   - Perform FFT on filtered signal to find the frequency with the largest amplitude
   - Convert this frequency to BPM to obtain RR_c

   1.2 **Generate RR_{list}:**
   - Sample respiratory rates in increments and decrements of 0.5 BPM around RR_c from $\min(\text{RR}_{list})$ to $\max(\text{RR}_{list})$:

$$\min(\text{RR}_{list}) = \max(7.5, RR_c - w/2)$$

$$\max(\text{RR}_{list}) = \min(42.5, RR_c + w/2)$$

   where w is the predefined length of RR_{list} and $7.5BPM$ and $42.5BPM$ are the smallest and largest human respiratory rate [177]

2. **Search for the best respiratory rate candidate (RR_{est}^{F})**
   2.1 **For each RR_{list}^{i} in RR_{list}:**
   - Set a BPF with cutoffs:

$$l^i = 0.65 \cdot \text{RR}_{list}^{i}/60, \quad h^i = 1.35 \cdot \text{RR}_{list}^{i}/60$$

   - Filter S_{HRV} with the BPF to obtain the breathing signal $Breath^i$.
   - Perform FFT on $Breath^i$ and find the frequency with the largest peak, f_{max}.
   - Convert f_{max} to beats per minute to estimate the respiratory rate RR_{est}^{i}.
   - Compute the difference $|\text{RR}_{est}^{i} - \text{RR}_{list}^{i}|$ and store it.

   2.2 **Min-max normalise the frequency difference list (F-Difference-list)**.
   2.3 **Select the three candidates with the local minima** as the best frequency-domain estimates (RR_{top}^{F}).

3. **Calibrate using time-domain analysis**
   3.1 **For each RR_{list}^{i} in RR_{list}:**
   - Estimate the respiratory rate in the time domain by counting the zero-crossing points from $Breath^i$.
   - Compute the time difference (T-Difference-list) and smooth it.

   3.2 **Sum the T-Difference and F-Difference values** for RR_{top}^{F} ($\text{RR}_{calibrated}^{TF}$)
   3.2 **Select the candidate with the smallest combined difference** from $\text{RR}_{calibrated}^{TF}$ as the final estimated respiratory rate, RR_{est}^{TF}.

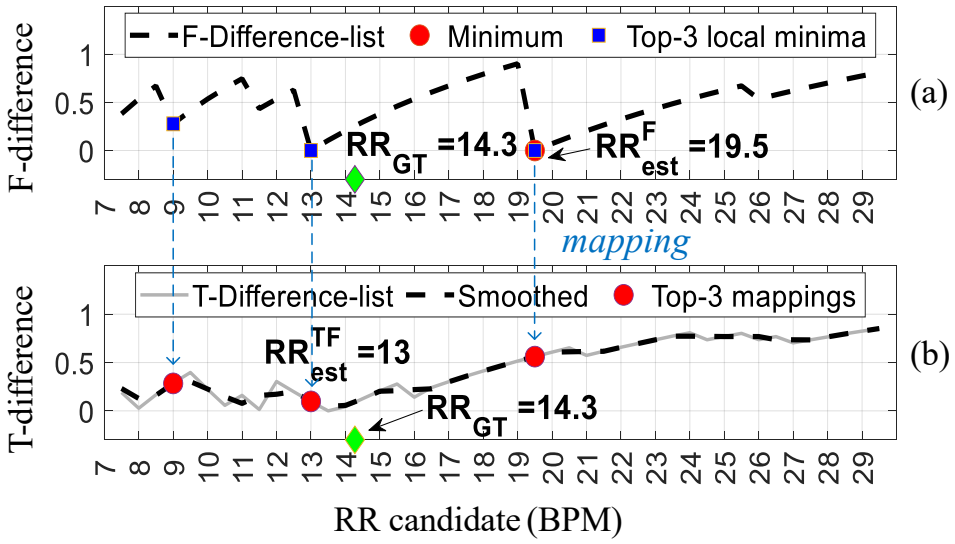


Figure 4.3.5: Breathing rate estimation. (a) Best respiratory rate candidate searching in the frequency domain. (b) Calibration from the time domain.

4.3.2 LRC-based respiratory rate monitoring

When rhythmic footsteps are present, we leverage the LRC-based respiratory rate monitoring pipeline for respiratory rate estimation.

Although LRC indicates a synchronisation between stride rhythm and respiratory rate, the LRC ratio between stride rhythm and respiratory rate is variable and unknown. Consequently, respiratory rate cannot be directly estimated from stride frequency alone. Inspired by previous studies [97, 98] which linked stride frequency with breathing signals to estimate the LRC ratio (but only under running scenarios with a fixed LRC ratio per window), we propose our pipeline. This pipeline addresses two unique challenges in achieving LRC-based respiratory rate estimation from in-ear audio:

- Respiration sounds are strongly interfered with by other sounds in the in-ear audio, especially footstep sounds which are strong and amplified due to the occlusion effect [163] (Figure 4.2.1).
- The LRC ratio varies within an estimation window, especially for walking and non-regular runners, as demonstrated in Figure 4.3.6. We analyse in-ear audio from two participants (User A, a non-regular runner, and User B, a regular runner) by segmenting their audio into 10-second intervals and calculating the mean LRC for

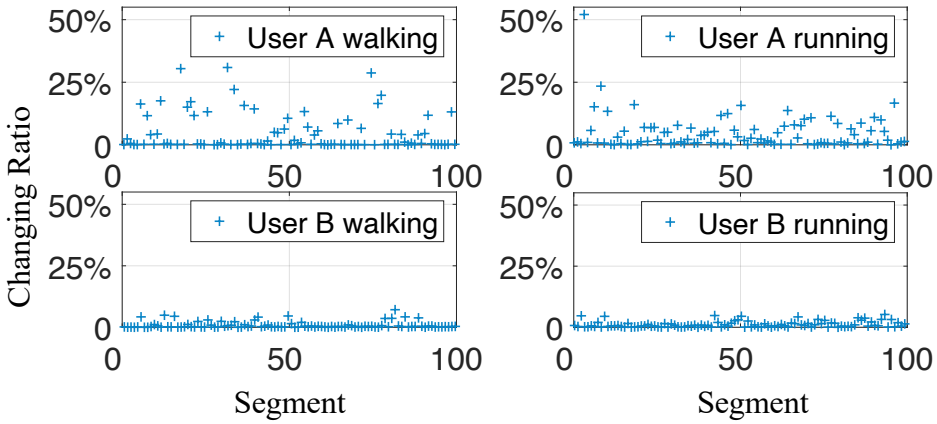


Figure 4.3.6: Changing ratio of ground truth LRC for user A and user B while walking and running, respectively. Each point represents the changing ratio between two adjacent segments, with higher ratios indicating higher irregularity.

each interval. We then compute the *changing ratio* of mean LRC values between adjacent segments to gauge irregularity. As depicted in Figure 4.3.6, this variability is especially evident in non-regular runners and during walking. This indicates that a constant LRC ratio cannot be assumed, meaning that our system must be able to accommodate changing LRC ratios within a single respiratory rate estimation.

We elaborate on our pipeline in the following sections.

Stride frequency estimation

We first detect the footsteps in the signal using the same approach as was used for heartbeat detection in Section 4.3.1. However, unlike with the heart signals, we filter with a 50Hz low pass filter, as in Chapter 3. By counting the number of detected footsteps, the stride frequency can be estimated.

Breathing extraction

To achieve LRC-based estimation, breathing sounds are required. In this component, we extract breathing-related features from the partly discernible breathing sounds as follows:

Pre-processing: Human breathing sounds typically fall within the range of 300Hz to 1800Hz [178]. Therefore, a bandpass filter with cutoff frequencies from 300Hz to 1800Hz is used on the input audio during light-intensity rhythmic footstep activities, such as walking.

For high-intensity rhythmic footstep activities, such as running, where step sounds severely overwhelm the breathing sounds in this frequency range, we use a bandpass filter with cutoffs of 2000Hz to 9000Hz to capture harmonics of breathing sounds in these higher frequencies.

Breathing template generation: We generate a breathing template (*i.e.,* signal features of a strong, clear breathing sound) to identify the probability of each frame containing breathing within the estimation window. To generate this template, we collected in-ear audio from a single user while sitting stationary and breathing loudly in a quiet environment. We conducted *Pre-processing* and performed *FFT feature generation* on it to generate the breathing template. We note that a uniform breathing template is used for all users in RespEar, without requiring individual calibration.

FFT feature generation: We divide the audio window into 40ms frames with a 20ms overlap and calculate the periodogram (which measures the power spectral density) of each frame [97]. Thereafter, we subdivide the breathing frequency range into 15 bins and sum the signal power in each bin from the periodogram. We therefore generate a feature vector with 15 features for each frame, one corresponding to each frequency bin. The breathing template is finally calculated by averaging the feature vectors of all frames.

Probability curve generation. For each estimation window, we perform *FFT feature generation* on all frames within it. For each feature vector (*i.e.,* corresponding to each frame), we calculate its similarity (S) with the breathing template using the cosine similarity [97]. Then, the probability of this frame containing breathing, $P(f)$, is computed as:

$$P(f) = \begin{cases} \frac{S-T}{1-T} & \text{if } S > T \\ 0 & \text{if } S \leq T \end{cases} \tag{4.1}$$

where T is a predefined threshold. The probabilities from all frames within the estimation window generate a breathing probability curve as shown in Figure 4.3.7(a).

Probability curve decomposition: Due to the low SNR from the strong interference from footstep sounds and light breathing sounds, the breathing pattern is overwhelmed by patterns of interference (Figure 4.3.7(a)). To remove the interference patterns, we decompose the probability curve into its constituent components using the SSA algorithm [160], as explained in Chapter 2. SSA is able to effectively separate the underlying components of the curve, allowing for the isolation of periodicity in order of significance, even within

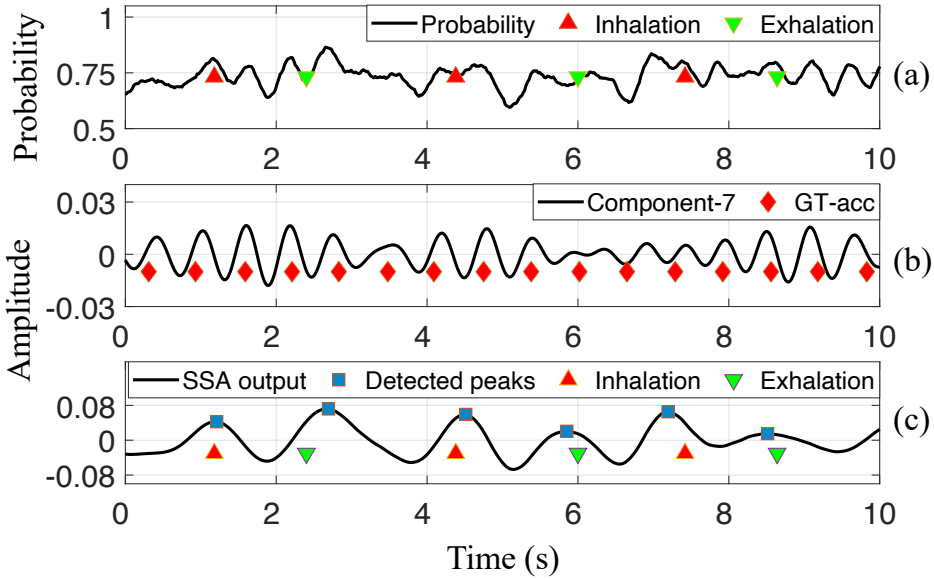


Figure 4.3.7: Breathing signal extraction. (a) Breathing probability curve. (b) One component of the curve related to steps while walking. (c) Extracted breathing pattern with peak detection to determine respiratory rate.

highly noisy time series data. Figure 4.3.7(b) shows component 7 of the decomposed probability curve which corresponds to the steps taken while the user is walking (*i.e.,* GT accelerometer data (GT-acc) in Figure 4.3.7(b)). Once the probability curve is decomposed into periodic components, we exclude the components related to interference and aggregate the breathing-related components for respiratory rate estimation, as described in the following section.

Breathing-related components aggregation

RespEar leverages a loose constraint which can adapt to changes in the LRC ratio to exclude components not related to respiration. Specifically, for each decomposed component of the probability curve, we count the number of peaks using peak detection. If the number of peaks falls outside the range of the minimum possible breathing rate (RR_{min}) to the maximum possible breathing rate (RR_{max}), the component is regarded as unrelated to respiration and removed.

Using the computed step frequency and the loose LRC constraint, RR_{min} and RR_{max} are

computed as:

$$\mathrm{RR}_{min} = \frac{\mathrm{SF}_{est} * N/fs}{\mathrm{LRC}_{max}} \tag{4.2}$$

$$\mathrm{RR}_{max} = \frac{\mathrm{SF}_{est} * N/fs}{\mathrm{LRC}_{min}} \tag{4.3}$$

where SF_{est}, fs and N are the estimated step frequency, sampling rate, and number of samples in the estimation window, respectively. LRC_{max} and LRC_{min} are the largest and smallest values of the LRC ratios in humans.

We use the LRC range of 1.9 to 4.9 for low-intensity rhythmic footstep activities [179], and 1.8 to 5.6 for high-intensity rhythmic footstep activities [180]. These ranges cover common LRC ratios in humans under each set of scenarios [179, 180], and fully cover the LRC ratios present in our collected dataset.

After excluding all breathing-unrelated components, we sum the remaining components into the extracted breathing pattern. Peak detection is then applied to this signal to estimate the final respiratory rate (Figure 4.3.7(c)).

4.3.3 Pipeline selector

The pipeline selector determines which processing pipeline should be selected for respiratory rate estimation, *i.e.,* RSA or LRC-based, based on the presence of either clear heartbeat sounds or footsteps in the input signal. If the LRC-based pipeline is selected, we further differentiate this into low-intensity and high-intensity rhythmic footstep activities, so that the correct algorithmic parameters can be applied to the pipeline.

We train our pipeline detector using support vector machines (SVM). The in-ear audio is split into 5-second segments and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from each segment and used as the input features to the SVM. 5-second segments were found to have the best tradeoff between temporal resolution for high-quality MFCC extraction and stability in capturing the transitions between states. This ensures that the classifier is responsive to changes in the state, while also reducing unreliable results due to short, noisy windows.

We use a two-stage classifier whereby first we classify a segment as sedentary (*i.e.,* strong presence of heart sounds) or active (*i.e.,* strong presence of rhythmic footstep sounds). If active, we further classify it into low-intensity and high-intensity rhythmic footstep

activities. Consequently, there are 12 detection results from one model during each 60-second estimation window, and we determine the scenario of the whole window through majority voting. Specifically, we empirically determine that only consistent results obtained for more than 75% of segments lead to reliable pipeline selection. Voting aims to handle transition windows between two states that could result in inappropriate pipeline selection. If there is no convergence, the window will be discarded. We implement the SVM on single-channel audio to reduce computational complexity.

4.4 Implementation

To collect data, we used the device detailed in Section 3.4.1 and recorded data from the in-ear microphone with a sampling rate of 22050 Hz. We use the Zephyr BioHarness 3.0 chest strap [65] to collect ground truth respiratory rate with a 25 Hz sampling rate. 18 participants (9 male and 9 female) took part in our data collection, which was approved by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge. The participants' ages ranged from 22 to 55 with an average age of 30±8. The data collection was conducted indoors in an office environment, with additional validation studies conducted outdoors. The participants' ages ranged between 22 and 51. The participants underwent sedentary and active activities, with each activity performed for 5 minutes. The activities, which are similar to those in Chapter 3 are:

1. Sedentary:
    (a) Sitting
    (b) Standing
    (c) Lying down
    (d) Listening to music (performed for the duration of one song)
    (e) Working in the wild
    (f) Uncontrolled cool down after exercise
2. Active:
    (a) Walking
    (b) Running

The activities encompass typical scenarios when a person uses earables and where respiratory rate monitoring would be desirable. Our data collection involved controlled activities and also in-the-wild scenarios to ensure the applicability of our methods to real-world use. No breathing rates were imposed, and participants were free to breathe as they wished. Breathing after exercise was performed immediately after the user had completed their running to capture their natural cool down breathing. While sitting and standing, users were asked to move their head three times to capture head motions to assess the impact of our interference artefact filtering algorithm. Active activities were performed on a treadmill and participants chose comfortable paces for walking and running. Overall, our dataset has an average respiratory rate of 18BPM±4.1BPM, with a range of 7BPM to 31BPM.

4.5 Evaluation

4.5.1 Metrics

We evaluate system performance using the MAE, MAPE and Bland-Altman plots as described in Section 3.5.1, where we compare respiratory rate estimations per window to the ground truth measurements for the corresponding window.

4.5.2 RespEar overall performance

Overall performance

We present the overall performance of RespEar in Figure 4.5.1(a, b). RespEar achieves an overall MAE of 1.71 BPM (MAPE of 9.68%), with a MAE of 1.48 BPM (9.12%) and 2.28 BPM (11.04%) for sedentary and active respectively. The Bland-Altman plot for the overall performance of RespEar is provided in Figure 4.5.1(c). The mean error is very close to zero (-0.02 BPM), indicating that RespEar does not systematically overestimate or underestimate respiratory rate. Additionally, the narrow limits of agreement, ranging from -4.8 to 4.76 BPM, demonstrate excellent agreement between RespEar and ground truth respiratory rate measurements, highlighting the system's accuracy. We also observe a uniform spread of data points across the range of the ground truth respiratory rate, indicating that the error remains consistent regardless of the ground truth respiratory rate. This further supports the reliability of our technique. Figure 4.5.1(d) provides the Bland-Altman plot of RespEar grouped by activity level. While there is a slight overestimation of respiratory rate under active conditions, the data points still exhibit a

largely uniform distribution. In contrast, the sedentary estimates show no bias and exhibit low error. Overall, our system demonstrates the ability to accurately and consistently estimate respiratory rates across various activity levels.

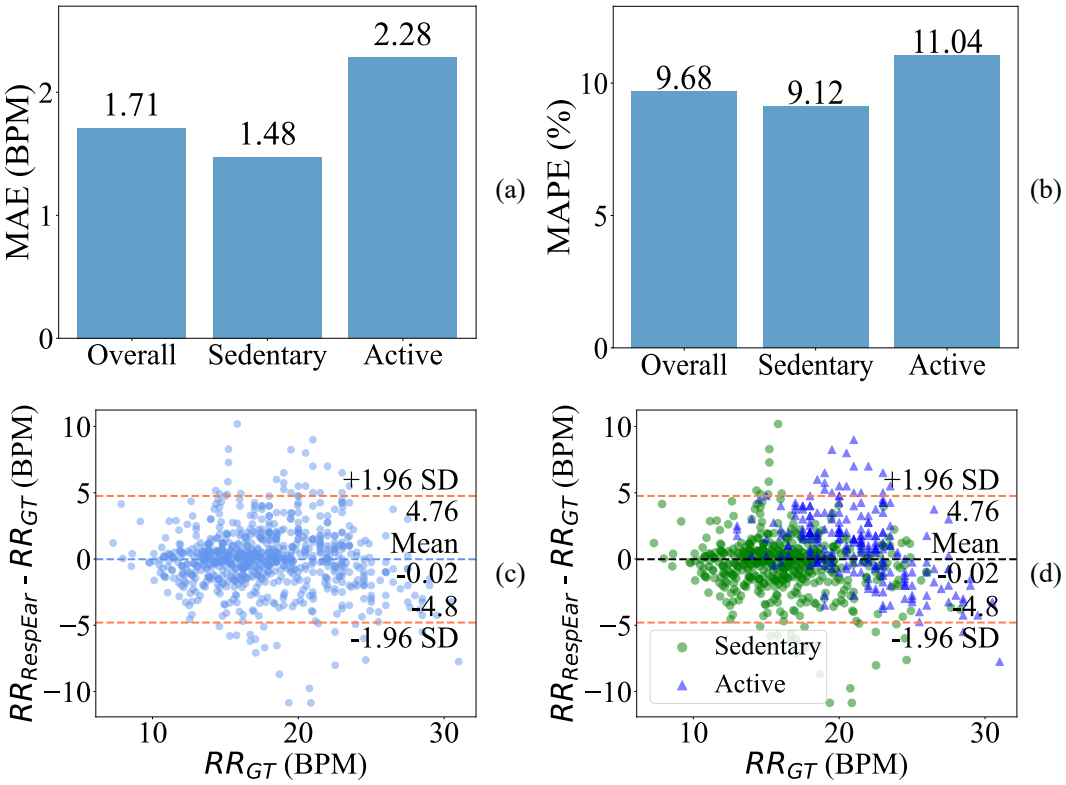


Figure 4.5.1: Overall system performance. Bar plot of (a) MAE and (b) MAPE. Bland-Altman plots of RespEar for (c) overall performance and (d) individual activity levels.

Performance per activity

Figure 4.5.2(a) provides a boxplot of the overall performance of RespEar for each activity. The performance of each sedentary activity is comparable. Slightly higher errors exist while listening to music (MAE=1.98 BPM) (detailed in Section 4.5.3), and working (1.56 BPM). This is because working is an uncontrolled activity and thus participants were more active during this task, leading to more interference artefacts. The estimation errors while walking and running are satisfactory where walking has a MAE of 1.75 BPM and a MAPE of 9.17%, and running has a MAE of 3.12 BPM with a MAPE of 14.01%. The slightly higher running errors are due to the increased interference from the footsteps.

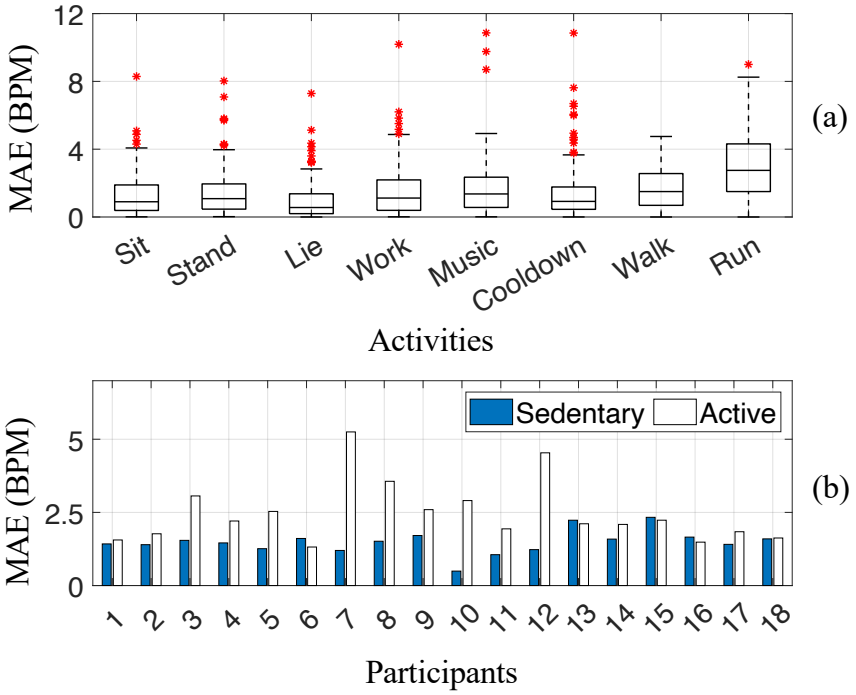


Figure 4.5.2: RR estimation errors for (a) different activities and (b) different participants.

Individualised performance

Figure 4.5.2(b) reports the overall performance of RespEar for each participant. It is evident that the MAE while sedentary is consistent amongst participants, with no participant exceeding an error of 2.3 BPM. There is much more variation amongst estimation errors while active: the smallest MAE is 1.26 BPM for participant 16, with the largest MAE being 5.25 BPM for participant 7. The majority of errors come from 2 participants, participants 7 and 12. Participant 7 ran at 5 KPH, which has a slightly worse performance than faster running speeds (discussed in Section 4.5.3). Participant 12's running generated large noise because their feet kept hitting the side of the treadmill, resulting in high-energy noise across all frequencies in the in-ear audio. However, regardless of this, the system still generalises well for the majority of participants over all activities.

Baseline comparison

We compare the performance of our system to that of existing works for each of the three sensors mentioned in Section 4.2.1. To perform this study, we collected data from 11 participants in a combination of indoor and outdoor settings, while the subjects were

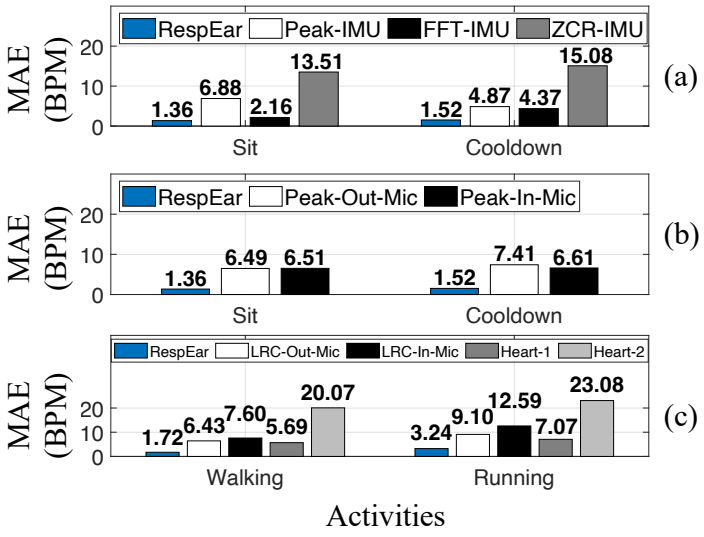


Figure 4.5.3: Accuracy comparisons for (a) IMU-based, and (b-c) audio-based approaches.

sitting still, cooling down after running (sedentary), walking and running (active), while wearing earable-based IMU, out-ear microphone and in-ear microphone. We implemented the three IMU-based algorithms for respiratory rate estimation under sedentary conditions in [94]: an FFT approach (FFT), a peak detection approach (Peak) and a zero-crossing rate (ZCR) approach (as shown in Figure 4.5.3(a)). These works were selected as they are the most commonly used in the literature for IMU-based respiratory rate detection while stationary. For the in-ear and out-ear microphones, we implemented the algorithm for sedentary estimation in [10] which uses a peak detection approach where peaks are detected from the envelope of the microphone signal (with the performance shown in Figure 4.5.3(b)). For active, we implemented the algorithm employed by [97,98] (LRC), and expanded upon it to calculate the respiratory rate using signals from the in-ear and out-ear microphones (Figure 4.5.3(c)). Through this comparison, we contextualise the performance of RespEar using the best-performing earable-based systems in the literature across all scenarios. Wrist-based systems were not implemented due to their vastly different signal properties and their limitation to stationary use (as discussed in Sections 2.4 and 4.1). Moreover, these systems employ algorithms similar to those used in earable IMU-based respiratory rate estimation, making a direct comparison less relevant.

Through this comparison, we contextualize the performance of RespEar against the best-performing earable-based systems in the literature across various scenarios. Wrist-based systems were not implemented due to their vastly different signal properties and their

limitation to stationary use (as discussed in Sections 2.4 and 4.1). Moreover, these systems employ algorithms similar to those used in earable IMU-based respiratory rate estimation, making a direct comparison less relevant.

From Figure 4.5.3, it is evident that under both sedentary and active scenarios, our system significantly outperforms the methods in the literature, thus highlighting both the strength of the in-ear microphone and our processing pipeline.

Using the algorithm in Chapter 3, we derived the HRV from the cleaned heart signal for the walking and running activities, and integrated it into the RSA-based respiratory rate monitoring pipeline (Heart-1). Following [46], we eliminated walking/running frequency components from in-ear audio and then applied a [0.1, 0.8] Hz band-pass filter for respiratory rate estimation (Heart-2). Figure 4.5.3(c) shows that although our hEARt algorithm has the best performance of the related work, all of these methods result in large estimation errors, showing that they cannot be applied for respiratory rate monitoring while active. This emphasises the need for our pipeline and highlights its superior performance.

4.5.3 Benchmark evaluations

Other active activities

We assessed RespEar's performance on a subset of 3 participants during various activities of different intensities that involve rhythmic footsteps, including step aerobics (StepA), climbing up and down stairs (StairUD), and rowing on an indoor rower (Row). We recorded 5 minutes of data per activity per participant, with activities performed in an uncontrolled manner. Figure 4.5.4(a) indicates that RespEar works properly during different rhythmic activities using the LRC-based pipeline, demonstrating RespEar's effectiveness for activities with rhythmic footsteps.

Outdoor performance

We also assessed system performance outdoors in an uncontrolled environment. The tests were performed on a concrete pavement outside an academic building next to a building site while active construction was occurring. We assessed the performance under different activities, including sitting, walking, running and cooling down. We recorded 5 minutes of data per activity for each participant. When walking and running, participants were free to select their preferred pace and move around the area. There were thus natural

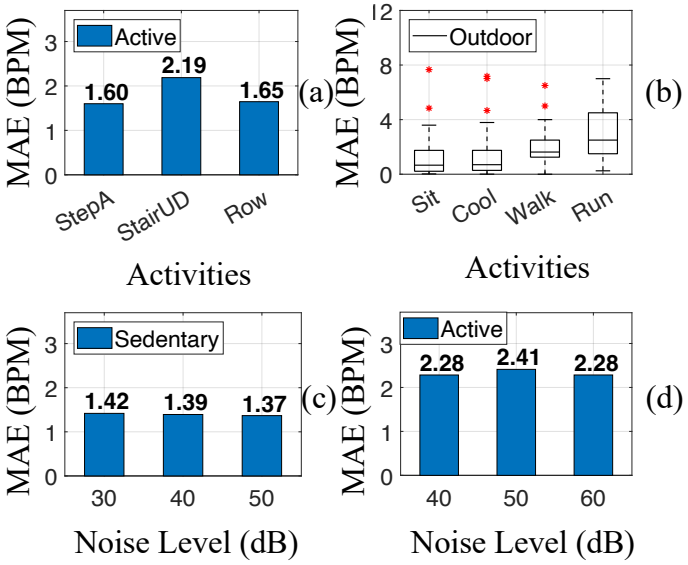


Figure 4.5.4: Estimation Errors (a) for other rhythmic activities, (b) while outdoors, (c-d) for different noise levels while sedentary and active respectively.

changes in pace throughout the experiment to test whether our system functions under both controlled and uncontrolled speeds. Figure 4.5.4(b) shows the results of this study, indicating that RespEar achieves robust performance outdoors, with results similar to that of the controlled, indoor study.

Impact of ambient noise level

Since RespEar is audio-based, it is essential to ensure that it functions as expected in the presence of ambient noise at different levels. We assess this performance in Figure 4.5.4(c,d). We see that over three ambient noise levels for sedentary (Figure 4.5.4(c)) and active (Figure 4.5.4(d)) activities, RespEar achieves consistent results. This is because by occluding the ear canal with earbuds, ambient noise is attenuated. In addition, the occlusion effect attenuates high-frequency external noise and amplifies low-frequency heartbeat and footstep sounds [162, 163], making our solution robust to noise variations.

Impact of the audio channel

Figure 4.5.5(a) provides the overall MAE in the left and right channels and the channel chosen by *automatic channel selection* (denoted by *two*) while sedentary. The individual

performance of the two channels is similar with a MAE of 1.8 BPM and 1.79 BPM for left and right respectively. After the *automatic channel selection*, the MAE is significantly reduced (1.42 BPM), indicating the efficacy of our design. Moreover, even from one channel, the performance is still good, proving that RespEar can be used even when the participant is wearing a single earbud. Figure 4.5.5(b) provides the MAE for the left, right and fusion of two channels (*i.e.,* we use the mean value of the estimated respiratory rate from two channels) while active. Again, we see that the fusion channel has a lower error than that of the two channels, with a MAE of 2.28 BPM for the fusion compared to 2.34 BPM and 2.56 BPM for left and right respectively.

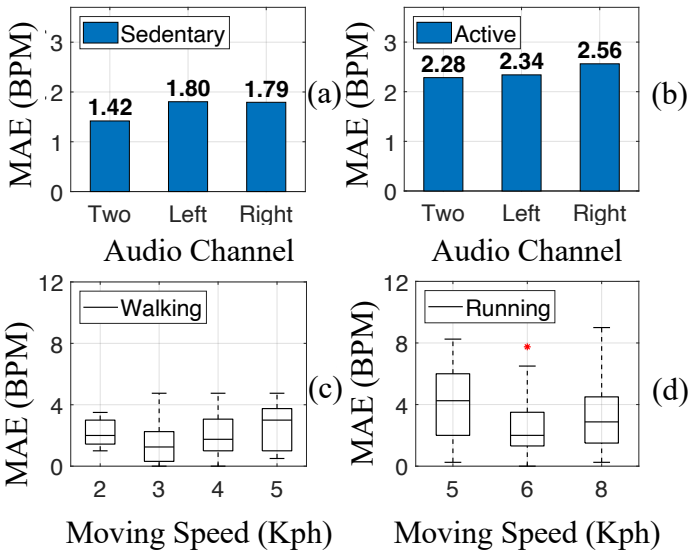


Figure 4.5.5: Estimation errors for (a-b) different audio channels and (c-d) different moving speeds.

Impact of moving speed

In Figure 4.5.5(c,d), we assess the impact of different speeds while active. For both walking and running, the median error remains similar regardless of the movement speed. We see a slightly higher error while running at 5 km/h. This may be attributed to the lower SNR of low-speed running, which causes similar interference from running footsteps but induces weaker breathing sounds compared to higher speeds. However, even with this slight trend, RespEar still achieves good results across different speeds.

Impact of music listening

We assessed the impact of listening to music on earables while participants were sedentary, walking and running. For each condition, we played music covering a range of common genres including pop, dance, alternative, classical and R&B at different volumes through the earable speaker on the left channel. Figure 4.5.6(a) and Figure 4.5.6(b) provide a comparison of performance per genre per activity and performance with different music volumes respectively. The performance across all genres and with different playing volumes while sedentary is comparable with the sedentary performance without music playing (the error of the left channel in Figure 4.5.5(a), *i.e.,* 1.8 BPM). This is due to the minimal overlap of heart sound frequencies with the music frequency, with only 0.5% of music energy lying in the 0-30 Hz range. While active, the error across all genres is slightly higher than the error without music. Specifically, errors are 2.50BPM and 2.92 BPM for soft and loud music respectively compared to 2.34 BPM when active without music (the error of the left channel in Figure 4.5.5(a)). Again, music does not significantly degrade respiratory rate estimation performance while active. There is again a non-overlap between music and stride frequency, meaning that music does not impact stride detection accuracy. Although music and breathing sounds have overlapping frequencies, we do not use breathing sounds directly but rather generate a breathing template to detect the probability of breathing sounds. This means that breathing and music can easily be differentiated, mostly mitigating the impact of music on accuracy.

In-the-wild performance

We asked two participants to wear the device for an hour in a busy office while undergoing standard daily activities. The participants worked at their desks, listened to music, walked around the office, and performed other activities as they wished, such as sipping coffee or using their cell phones. Figure 4.5.7 shows the tracking performance for two participants, *i.e.,* the continuously estimated respiratory rate from RespEar (one respiratory rate estimation per 30s) compared with the ground truth respiratory rate (GT_{RR}). Participant A (shown in Figure 4.5.7(a)) worked and walked around the office. Participant B (Figure 4.5.7(b)) worked at their desk and listened to music while working. Participant A achieved a MAE of 0.97 BPM and 1.83 BPM while working and walking respectively, and Participant B achieved errors of 1.01 BPM and 0.58 BPM while working and listening to music respectively. These errors are consistent with the results obtained in the laboratory study for these two participants, proving that RespEar has excellent performance both in controlled laboratory

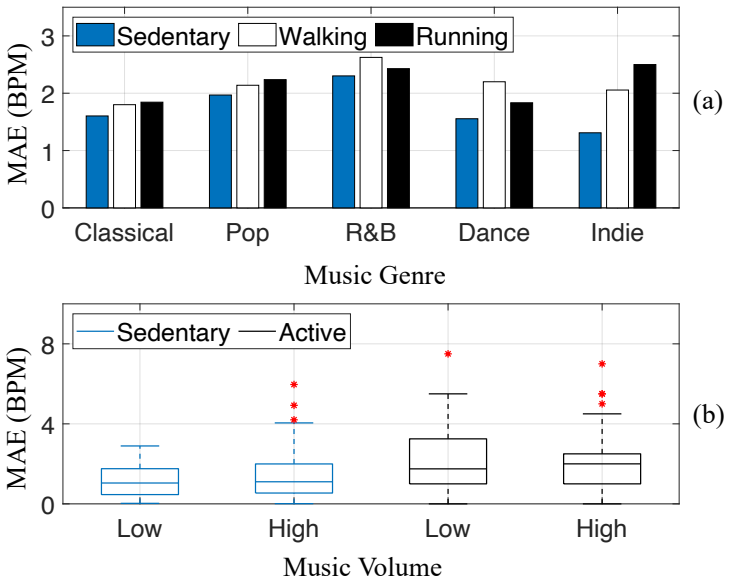


Figure 4.5.6: Errors with music (a) across genres and (b) at different volume levels.

settings and uncontrolled, real-world settings. It is also clear that RespEar can accurately track respiratory rate longitudinally, even in an uncontrolled setting.

4.5.4 System components evaluation

Pipeline selector

Figure 4.5.8(a) provides the results of our pipeline selector module using SVM on 5 s segments. The SVM is trained on 13 participants' data (randomly chosen during training) and tested on the remaining 5 participants to ensure participant independence of the train and test sets. We implemented 5-fold cross-validation and report the average results over 5 folds. Our system is able to select pipelines with excellent performance, achieving 100% accuracy for determining whether a window is active or sedentary (SVM-Stage-I), and 99% accuracy for determining whether an active window is low-intensity or high-intensity rhythmic footstep activities (SVM-Stage-II). With majority voting of results across an estimation window, the detection accuracy on both tasks is 100%.

Accuracy of HRV estimation and stride detection

To assess the accuracy of our HRV estimation, we compute the MAPE between the ground truth HRV from the ECG chest strap and our estimated HRV on beat-to-beat

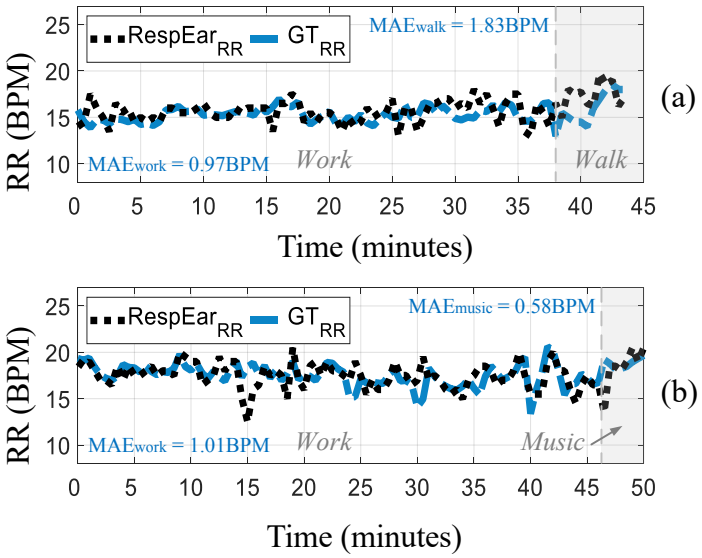


Figure 4.5.7: Errors of longitudinal in-the-wild tracking for (a) participant A and (b) participant B.

basis (Figure 4.5.8(b)), where we compare performance on the left channel, right channel and *automatic channel selection*. Using *automatic channel selection*, we achieve the best performance with a MAPE of 3%, which is competitive with reported results for ECG on well-known datasets, *e.g.,* the MIT-BIH Arrhythmia Database [181]. Our results for stride frequency estimation are provided in Figure 4.5.8(c). Our system detects strides with a MAPE of less than 3% for both walking and running using channel fusion, again competitive with literature on in-ear step counting [163].

Ablation study

In this section, we assess the impact of some of the design choices in RespEar.

1) Interference artefact filtering: In Figure 4.5.9(a), we examine the impact of our interference artefact filtering. The MAE without interference artefact filtering is 1.63 BPM, and that with interference artefact filtering is 1.45 BPM, showing that we achieve a 12% decrease in error. We also see that interference artefact filtering decreases variance and median estimation error, resulting in more robust estimations.

2) Adaptive breathing signal extraction: Figure 4.5.9(b) illustrates the performance of RespEar while sedentary (using in-ear audio) compared to the performance obtained when directly using the adaptive filter generated from the ground truth respiratory rate

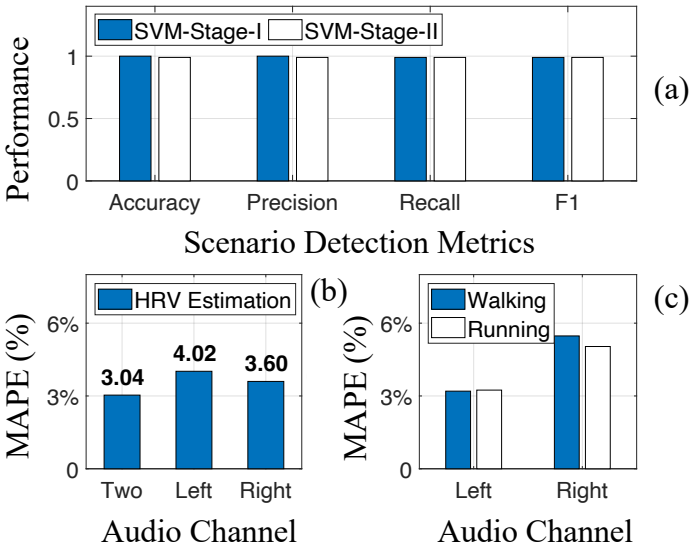


Figure 4.5.8: Performance of system components. (a) Pipeline selector, (b) HRV, and (c) Stride frequency.

(from ECG as detailed in Section 4.3.1). RespEar achieve a MAE of 1.48 BPM, and with the adaptive filter using the ground truth respiratory rate, we obtain an error of 1.44 BPM. We thus see that the proposed design leads to estimations that are very close to the upper bound of performance that is possible using the RSA-based respiratory rate estimation principle.

3) Calibration from T-difference-list: Figure 4.5.9(c) illustrates the importance of our calibration technique which fuses the T-Difference-List and the F-Difference-List. By incorporating calibration with the T-difference list, we achieve a 24% decrease in error over using the T-difference list in isolation, and a 32% decrease in error over only using the F-difference list. Thus, our calibration technique results in a significant performance improvement.

4) FFT feature generation: In Figure 4.5.9(d), we compare our FFT features with MFCC features for creating the breathing features template for respiratory rate estimation during active scenarios. Although MFCCs are most commonly used for audio feature extraction [182], we gain a significant performance increase through the use of FFT features for both walking and running.

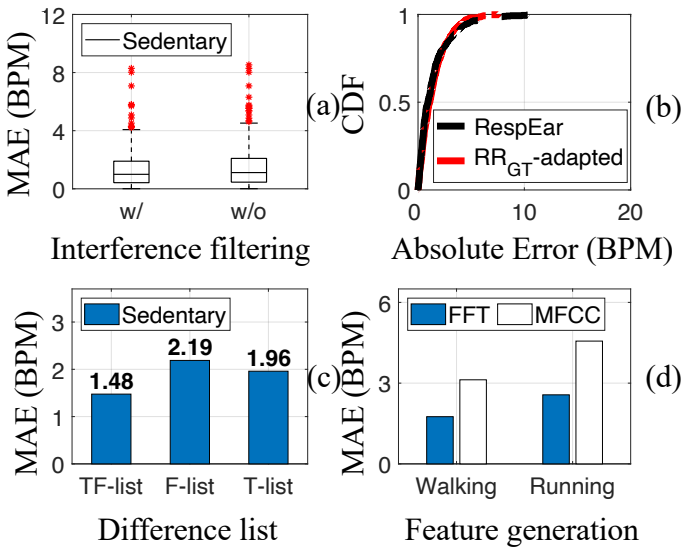


Figure 4.5.9: Impact of (a) interference filtering, (b) adaptive BPF, (c) difference list, (d) FFT features.

4.5.5 System overhead on a smartphone

To create a portable solution, we deployed RespEar on an iPhone 12 Pro with 8 GB of memory, and a 2477 mAH battery capacity (2815 mAH battery with 88% battery health) to measure system overhead. When run on an iPhone 12 Pro, RespEar's latency is 3.11 s per window while sedentary and 12.27 s per window while active. Since processing occurs in 60-second windows, with a 30-second overlap, a new respiratory rate estimate can be made every 30 seconds under both conditions, implying that our system can run in real-time. We have not considered possible data transfer costs via radio frequency (*e.g.,* BLE) which would add additional small delays depending on the scenarios. When run continuously for an hour, RespEar consumed 4% and 14% battery for the RSA-based and LRC-based pipelines, respectively. The LRC-based pipeline consumes more power due to the longer latency on account of the more complex algorithm. To contextualise this, if playing music for an hour, the same phone battery decreased by 8%, showing that our application lies within standard levels for battery consumption. However, to reduce this, further standard processing optimisations could be applied. In terms of memory, RespEar consumes a maximum of 49.1 MB and 49.3 MB per window equating to 0.3% of the available memory on the device. Overall, we see that RespEar can feasibly be run on a smartphone longitudinally to potentially provide real-time respiratory rate monitoring from earables.

4.6 Conclusion

This chapter presented RespEar, the first earable system for continuous, non-invasive, and reliable respiratory rate monitoring across both sedentary and active conditions. RespEar employs in-ear microphones and leverages unique relationships between our cardiovascular, gait and respiratory systems to optimise respiratory rate detection. Using our earable prototype, we implemented RespEar and conducted extensive experiments to evaluate its performance. The results demonstrate that RespEar outperforms the state-of-art, and is robust in a variety of contexts and activity intensities.

This chapter therefore demonstrated the potential of using in-ear audio to monitor breathing rate under a variety of activities of daily life. We built upon the findings in Chapter 3 to leverage various physiological couplings based on the sounds of footsteps and heartbeats to determine breathing rate. In the next chapter, we will further examine the heart signals captured inside the ear canal and investigate the feasibility of monitoring stroke volume using in-ear audio.

Chapter 5

Stroke Volume Monitoring

5.1 Introduction

In Chapters 3 and 4, we showed the possibility of capturing clear heart signals using in-ear audio. In this chapter, we build upon this knowledge and the research gap in Section 2.5 to explore, for the first time, the feasibility of stroke volume monitoring using in-ear audio.

As described in Chapter 2, stroke volume, the volume of blood ejected by the left ventricle in each contraction, is one of the fundamental measures of cardiovascular physiology. It is closely related to the cardiac output, which represents the total volume of blood pumped by the heart per minute. Through its connection to cardiac output, stroke volume provides an estimate of the body's ability to deliver oxygen to the tissues (tissue perfusion) and its ventricular function [183], making it crucial in evaluating the health of the circulatory system.

Stroke volume is a key parameter in sports medicine and exercise physiology, used to assess overall cardiovascular health and fitness. Higher stroke volumes are often observed in athletes with more efficient cardiovascular systems, making it an excellent indicator of cardiovascular fitness [184, 185]. There is thus potential for stroke volume measurements to be used to monitor improvements in cardiovascular fitness over time due to exercise and to customise training programs to meet lifestyle and fitness goals. Moreover, stroke volume directly reflects the efficiency of the myocardium, the resistance to blood flow from the heart (afterload), and the return of blood to the heart (preload). When one of these is disrupted, the resulting stroke volume changes can serve as an early indicator of subclinical changes in

cardiovascular function, potentially revealing abnormalities in healthy individuals before observable disease symptoms emerge. Stroke volume therefore has the potential to be used for prediction, early diagnosis and continuous monitoring of cardiovascular health in normal individuals. Clinically, low stroke volume can be a strong indicator of heart failure or other cardiac conditions, such as arrhythmias and heart valve diseases [103, 186]. Therefore, understanding and measuring stroke volume is essential for evaluating cardiac function, which plays a critical role in diagnosing and managing cardiovascular diseases [103] and monitoring the progression of disease states in patients with cardiac pathologies.

As outlined in Chapter 2, the gold standard for stroke volume measurement is the Direct Fick method [104, 105], however, this method is highly invasive, requiring multiple catheters, and very labour-intensive [106]. As a result, echocardiography (or cardiac ultrasound) has become more prevalent as a non-invasive alternative for calculating stroke volume [105]. However, this technique requires highly skilled operators to accurately determine stroke volume, limiting its widespread use [105]. Additionally, it is not suitable for continuous stroke volume monitoring. For non-invasive, continuous stroke volume measurements, finger-cuff systems, such as the Finapres NOVA [107], can be used. While reliable, these systems are cumbersome and require calibration using a stroke volume measurement from an echocardiogram [109], limiting their utility. In summary, all of these methods only have clinical use due to their dependence on complex procedures, highly specialised and expensive equipment, and trained personnel. Additionally, these techniques have limited accessibility worldwide due to their high cost and limited scalability.

Wearable devices have emerged as an alternative for non-invasive stroke volume estimation. Chest patches equipped with a single-lead ECG and a triaxial accelerometer [13] or with a PPG sensor [14] have been used for cardiac output and stroke volume measurements. However, these systems have a conspicuous and uncomfortable form factor and often require calibration with a blood pressure cuff. Finger cuffs equipped with PPG have also been used [15], though their bulky form factor limits real-world applicability. In addition, a bathroom weighing scale using ballistocardiography, ECG, and impedance plethysmography has been developed to estimate stroke volume [104]. However, this device relies on specialised sensors not found in standard scales, requiring users to purchase a new, costly device for monitoring. Additionally, it relies heavily on user adherence, limiting monitoring to the specific moments when the user steps on the scale, a well-known challenge with medical devices [114]. *Thus, there is currently no solution that provides stroke volume monitoring in a convenient, unobtrusive, and widely accessible form factor using only commodity devices*

and their native sensors.

Earables present an opportunity to achieve this through their natural integration into people's lives and daily routines, thus enabling passive sensing of stroke volume, without relying on user adherence. The cardiac signals that can be captured using in-ear microphones (Chapters 3 and 4) contain fundamental information about events in the cardiac cycle (such as valve closures, contractions, and electrical stimulations) [60], which are related to stroke volume [13, 109]. Additionally, studies have demonstrated a correlation between the amplitude of heart sounds collected on the chest and cardiac output, and by extension, stroke volume [111]. Therefore, there is the potential to use in-ear heart sounds to infer stroke volume, presenting a unique opportunity to develop a commodity wearable-based method for stroke volume monitoring outside clinical settings.

In this chapter, we demonstrate for the first time the feasibility of estimating stroke volume using sensors on commodity earables in a robust, generalisable way. By capturing cardiac signals from inside the ear canal through sensors embedded in commodity earbuds, we open the door to affordable and scalable stroke volume estimation. To achieve this, we encounter three main challenges. Firstly, while relationships between various heart signal modalities and stroke volume have been identified, in-ear heart signals are complex and mapping them to stroke volume remains unclear and challenging. Secondly, as is common when developing new wearable technologies, collecting a large-size dataset is challenging due to the requirement for equipment and expertise, which hinders the development of high-performing algorithms. Finally, a key challenge in machine learning research is ensuring model generalisability, particularly when faced with data from unseen subjects with varying stroke volumes, and in-ear audio signal properties.

As a solution, we develop a two-stage training process that combines generative self-supervised pre-training with fine-tuning. We introduce CardiAural, a pipeline designed to estimate stroke volume using in-ear audio. First, to address the initial challenge, we apply self-supervised learning techniques that mask and reconstruct intricate in-ear heart signals. Training the model to reconstruct randomly masked spectrogram representations allows it to learn complex and robust features crucial for accurate stroke volume estimation. Second, to tackle the issue of data scarcity, we utilise an in-house dataset of unlabelled earable data for self-supervised pre-training. This enables the model to acquire generalised and robust representations of in-ear audio without relying on large labelled datasets. Finally, we fine-tune the model with a smaller, highly specialised dataset specifically collected for

stroke volume estimation. This two-stage approach not only mitigates data scarcity but also enhances the model's ability to generalise effectively.

To evaluate the performance of CardiAural, we collected in-ear audio using the custom earbud device introduced in Chapter 3. We compare our stroke volume estimations to those obtained with a clinically validated device in a cohort of 23 healthy participants with average weight ($\bar{x} \pm \sigma$) of 76±15 kg and average height of 158±9 cm, and assess performance before and after isometric exercise. Overall we achieve a MAPE of 6.83% in estimating the average stroke volume of unseen subjects, by training the model on data from all other participants. This demonstrates the model's strong ability to accurately predict stroke volume, even for unseen participants. Additionally, we achieve a percentage error in the limits of agreement of 11.05% which is well below the 30% error considered acceptable for a new stroke volume measurement device [187], effectively indicating that commodity earables can be a suitable tool for stroke volume measurements. More specifically, our findings highlight the feasibility of using a single sensor on a commodity earbud to obtain accurate stroke volume measurements, paving the way for out-of-the-clinic cardiac monitoring using wearable devices.

The contributions of this chapter can be summarised as follows:

- We explore stroke volume estimation with in-ear microphones and present an investigation of the relationship between common audio features and demographic factors with stroke volume.

- We propose a novel pipeline for average stroke volume estimation that uses a generative self-supervised learning framework with a masked autoencoder. To enhance model generalisability to new users, we use transfer learning to pre-train the model with a large dataset of unlabelled in-ear audio data.

- We collected data from 23 participants using our earbud prototype. Results show that we can estimate stroke volume with a MAE of 5.27 mL (6.81%) and 5.21 mL (6.85%) before and after isometric exercise respectively. The predicted stroke volumes are also highly correlated to the ground truth with a Pearson correlation coefficient of 0.94, proving the effectiveness of our system.

The techniques and results in this chapter have been presented in [19].

5.2 Primer

In this section, we discuss in-ear heart signals and then examine the relationship between features of these signals and stroke volume.

5.2.1 In-ear heart signals

As discussed in Chapter 2, when the ear canal is blocked, the occlusion effect occurs [16, 60]. This effect can be leveraged to capture heart signals inside the ear canal, as proven in Chapters 3 and 4. Early works [16, 60] (including the work presented in Chapter 3) viewed these signals as heart sounds, attributed to the turbulent flow of blood in nearby vasculature. However, recent studies [188] suggest that these signals are more closely related to changes in local pressure due to pressure waves from heartbeats. Regardless, all literature agrees that these signals relate to heart activity as validated in Figure 5.2.1 and in Chapter 3.

While the precise origin of these in-ear heart signals - whether heart sounds or pressure waves - remains unclear, cardiac signals of various kinds have been successfully used to estimate stroke volume, which motivates our study. For example, previous research [104] has utilised features from ballistocardiography on a weighing scale, while other studies [13] have used features of seismocardiography signals on a chest patch for stroke volume determination. Both types of sensors are sensitive to body vibrations caused by heart activity, similar to the in-ear signals which are related to ear canal vibrations. Furthermore, research leveraging heart sound features [109, 111] collected using stethoscopes, and studies using arterial pressure waveforms [189] have demonstrated success in estimating stroke volume. Therefore, regardless of whether the in-ear signals are heart sounds or pressure waves, their strong correlation with cardiac activity suggests that they could be effectively used for stroke volume estimation.

5.2.2 Correlation between in-ear heart signals and stroke volume

We provide an example of the ground truth stroke volume collected from two participants and their corresponding in-ear audio in Figure 5.2.1. The corresponding ECG signals are also collected to validate the effectiveness of in-ear audio for capturing cardiac activity. Figure 5.2.1(a,b) present the stroke volume for one participant before and after isometric exercise respectively, and (c,d) present the same for a second participant. For each participant, 60 seconds of data per condition is shown. For Participant 1, the mean stroke

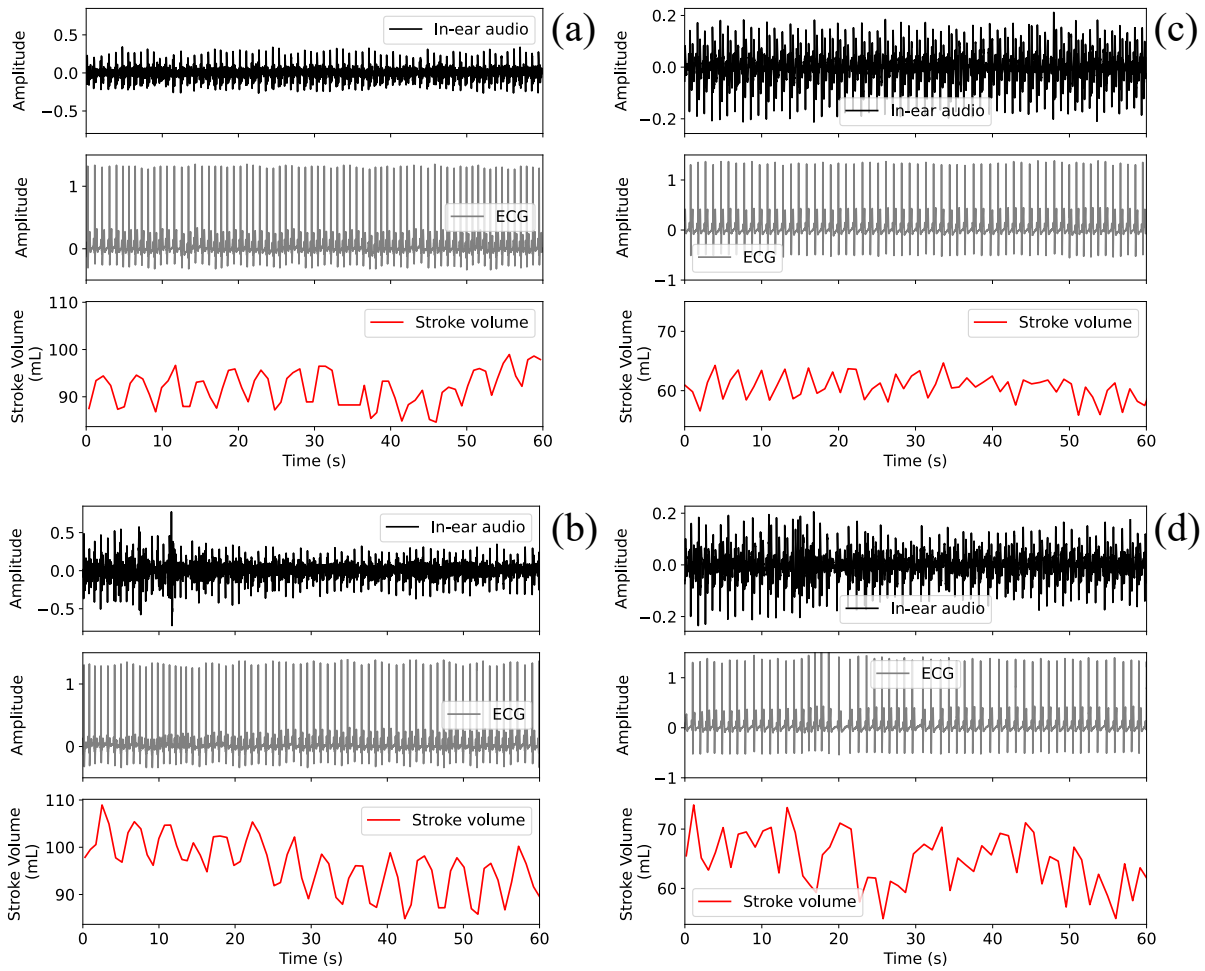


Figure 5.2.1: 60 seconds of in-ear audio, ECG and stroke volume for two participants before and after exercise. (a)-(b) Participant 1 before and after exercise respectively, (c)-(d) Participant 2 before and after exercise respectively.

volume in the 60 seconds before exercise was 91 mL and 96 mL afterwards, with an average heart rate of 70 BPM. For Participant 2, the mean stroke volumes were 60 mL and 65 mL, respectively, with a heart rate of 61 BPM. Although both participants are the same age and sex, their differing fitness levels (based on their exercise frequency) are reflected in the differences in stroke volume.

Furthermore, the alignment of the peaks in the in-ear heart signals with the ECG signals demonstrates their correlation, providing evidence that variations in the in-ear signals are due to heart activity. Additionally, the comparison of in-ear audio signals with stroke volume signals in Figure 5.2.1(b,d) reveals that changes in the in-ear audio correspond to

changes in stroke volume in the post-exercise condition, where stroke volume decreases as the participant returns to a resting state. Conversely, immediately after the isometric exercise, both participants showed an increase in the amplitude of their in-ear audio signals compared to their resting state, which corresponded to elevated stroke volumes. In the pre-exercise condition, participants maintained their baseline stroke volume, resulting in minimal changes and stable waveforms in the in-ear audio signals.

We further extracted common features of audio signals [190] including signal statistics and time and frequency domain features, as well as heart signal features such as inter-beat interval (IBI), heart rate, and the amplitude of the peak of the in-ear heart signal for each heartbeat. We then computed the Pearson correlation between the audio features in a 10-second window and the stroke volume for that window, removed any features with a correlation of less than ± 0.1, and plotted the results. In Figure 5.2.2, we explore the resultant relationship between stroke volume and in-ear audio features.

There is a moderate negative correlation between stroke volume and both the amplitude of the largest peak in the in-ear audio signal per heartbeat and heart rate (heart-related features). There is also a moderate positive correlation between the stroke volume and the skewness, zero crossing rate (ZCR), and the fourth MFCC component [190] (audio-related features). Other features exhibit weak positive and negative correlations. In summary, weak correlations exist between stroke volume and both audio and heart-related features, showing that in-ear heart signals have the potential to be used to estimate stroke volume. However, the findings indicate that handcrafted features are not sufficient for achieving good results that are generalisable to differing signal properties. Deep learning models are well-known for their ability to approximate complex relationships and identify intricate patterns and features from data, making them a promising solution to address this challenge [191].

Additionally, the lack of generalisability in the extracted features is clear when comparing the in-ear audio signals in Figure 5.2.1(a-d). It is evident that the amplitude range of the in-ear audio of the two participants differs and the signals present distinct properties per person. Specifically, comparing common metrics of signal characteristics, the root mean square energy (RMSE) of the in-ear audio signal from Participant 1 is 19.98, with a skewness of 0.28, kurtosis of 2.38 and ZCR of 0.025. In contrast, for Participant 2, the RMSE is 11.30, with a skewness of 0.13, kurtosis of 0.72 and ZCR of 0.019. This substantial variability between participants weakens the correlations shown in Figure 5.2.2 and poses challenges for generalising the findings to different users. Therefore, it is crucial to develop

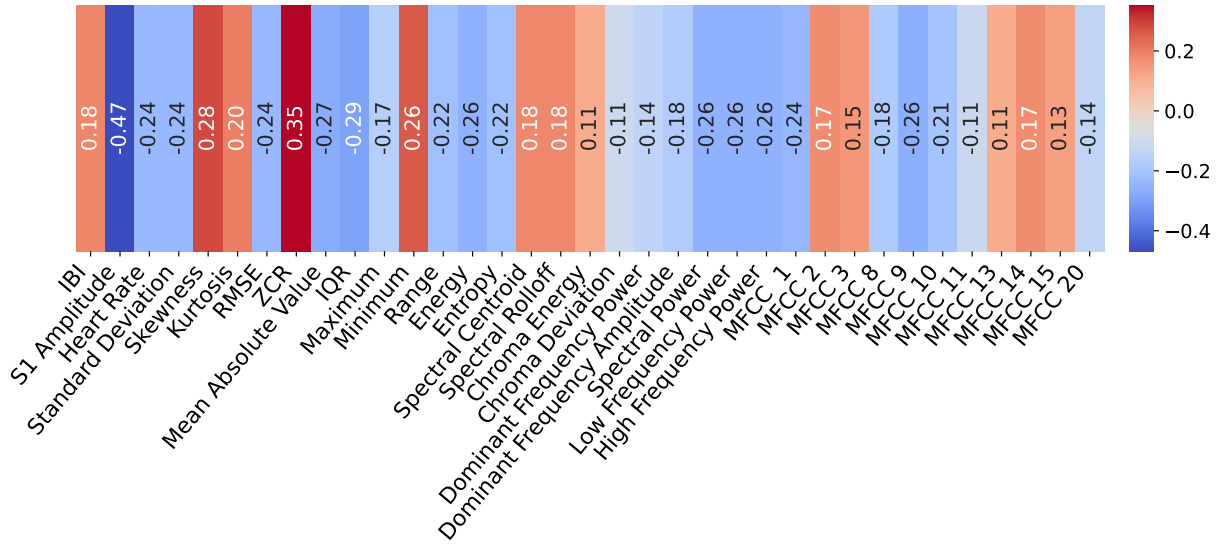


Figure 5.2.2: Correlations between features extracted from the in-ear audio and stroke volume. Weak correlations exist between heart-related properties and stroke volume, and between audio features and stroke volume. IBI: Interbeat-interval. RMSE: root mean square error. ZCR: zero crossing rate. IQR: interquartile range. MFCC: mel frequency cepstral coefficient.

techniques that can adapt to this inter-participant signal variability.

5.3 System Design

To address the challenge of generalisability and the need for advanced feature extraction uncovered in Section 5.2, we present our overall system pipeline in Figure 5.3.1. Our system takes as input in-ear audio containing heartbeat sounds and predicts the stroke volume for each window to ultimately compute the average stroke volume of a user. To achieve this, we use a generative self-supervised learning framework with a masked autoencoder. In the following sections, we describe the processing pipeline in detail.

5.3.1 Pre-processing

We first downsample our collected audio to 600 Hz to reduce computation complexity and divide the audio into 10-second segments with an 8-second overlap between successive segments. To remove unrelated signal components, we apply a 100 Hz low-pass filter. Similar to Chapter 3, we compute a mel spectrogram [132] on our 10-second windows using 32 mel

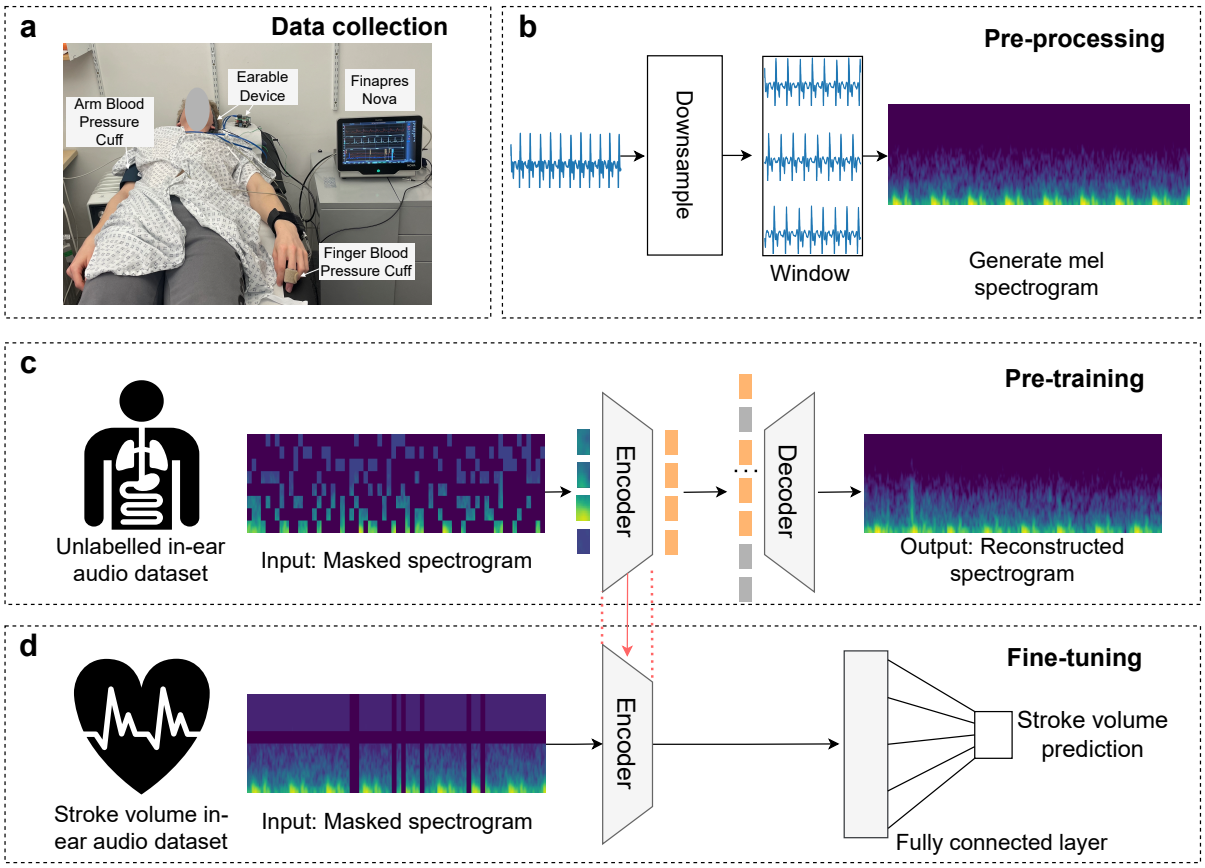


Figure 5.3.1: System processing pipeline. (a) Data collection protocol whereby data is simultaneously collected from our earable prototype and the ground truth Finapres NOVA continuous stroke volume measurement device. Not pictured is the isometric exercise where the participant squeezes a stress ball. (b) The pre-processing pipeline for the audio signals involves downsampling, segmenting the audio into 10-second windows with 8-second overlap, and generating a spectrogram for each window. (c) An unlabelled dataset of in-ear audio is used to pre-train the encoder-decoder network with masked spectrograms, employing a random, unstructured masking strategy. (d) The pretrained encoder is fine-tuned using the stroke volume in-ear audio dataset to predict stroke volume per 10-second window. During fine-tuning, a structured 2D masking strategy is used.

bins to accentuate low-frequency heart activity. We then interpolate this to 128 bins to match the size of the AudioMAE input [132]. As illustrated in Figure 5.3.1(b), the output of this pipeline is a 128×1024 mel spectrogram for each 10-second window.

5.3.2 Deep learning architecture

Figure 5.3.1(c,d) provides an overview of the pipeline used in CardiAural. Motivated by the recent success of vision transformers in audio classification, which effectively capture intricate spatial and temporal patterns in spectrogram data using self-attention mechanisms [192,193], we apply vision transformers (introduced in Chapter 2) to in-ear audio spectrograms. This approach enables the model to learn complex and robust features essential for accurate stroke volume estimation.

To achieve this, we use generative self-supervised learning, implemented as a masking and reconstruction task on the input spectrograms. Self-supervised learning (as described in Chapter 2) is a form of unsupervised learning that uses self-generated challenges to create supervised tasks from unlabelled data [194]. In generative self-supervised learning, the challenge is to reconstruct the input sample, where the sample itself supervises the training of its reconstruction [195]. In our system, this involves spectrogram reconstruction, where the model predicts masked portions of a spectrogram using their unmasked counterpart as the reference label. This approach enables the model to learn meaningful and robust data representations from the intricate in-ear heart signals, thus enhancing its ability to generalise to new users and varying signal properties [194].

Specifically, our pipeline uses the Audio-MAE model (as described in Chapter 2), which is an extension of a masked autoencoder from the image domain to audio [132]. At a high level, the network consists of a Transformer encoder and decoder. The network input is masked spectrograms (spectrograms with random binary patches applied to them) of audio and the network is only trained on the visible, unmasked patches. In doing so, the encoder extracts useful features from the masked spectrogram, enabling the decoder to reconstruct the original spectrogram. The encoder and decoder use a stack of 12-layer ViT-Base Transformers [196]. However, the decoder enhances the basic transformer module to incorporate local attention so that the model takes into account the position and order of the spectrogram features, both of which are key to the semantics of the sound itself.

5.3.3 Pre-training

Two of the key challenges in this work are the scarcity of labelled data and the need to enable the system to generalise to unseen subjects (as shown in Section 5.2.2), thereby allowing new users to achieve accurate stroke volume estimations without the need for expensive labelled data collection. To address these challenges, we implement a two-stage training process that combines generative self-supervised pre-training with fine-tuning.

Specifically, the model is initialised using weights obtained by training on the AudioSet dataset [132,197]. AudioSet is an audio dataset containing approximately 2 million 10-second audio files from YouTube videos [197]. This dataset is highly diverse with 527 classes ranging from music and vehicle noises to a small subset of respiratory sounds, including snoring and breathing. While AudioSet provides useful general audio representations, it is highly non-specific and not tailored for in-ear microphone data, which has significantly different signal properties. For example, music data may include frequencies up to 16 kHz, whereas the highest frequency of interest for in-ear cardiac signals is around 50 Hz [162, 163]. Therefore, in-ear audio requires a focus on lower frequency bands that are not well represented in general audio datasets. To better adapt the model to our specific task, we pre-train it using self-supervised learning on a dataset of unlabelled in-ear audio. The Earable dataset, detailed in Table 5.3.1, contains in-ear audio collected during stationary or mostly stationary activities, such as lying down, sitting, working at a desk, or speaking. In total, we pre-trained our with using 34940 10-second audio samples from this dataset.

During pre-training, we apply unstructured masking [132], where random pixels are removed, as shown in Figure 5.3.1(c). We use a masking ratio of 0.7, meaning that 70% of the pixels in the spectrogram are masked. The network was trained for 300 epochs with a learning rate of 1×10^{-3} and a batch size of 64, using mean squared error (MSE) loss. The final encoder from this pre-training phase is then used in the fine-tuning step.

5.3.4 Fine-tuning

The process of fine-tuning is illustrated in Figure 5.3.1(d). During fine-tuning, we remove the decoder layers from the pre-trained model and retain only the encoder with a single fully connected layer on top for stroke volume predictions. We fine-tune this model with the smaller, highly specialised dataset specifically collected for stroke volume estimation. For fine-tuning, we apply 2D masking as shown in Figure 5.3.1(d) [132] where we randomly mask bands of the spectrogram both in frequency and in time, with a masking ratio of 0.1.

Table 5.3.1: Earable dataset.

Dataset Name	Description	Duration (minutes)
hEARt [162]	Heart rate estimation dataset encompassing 2 activities with 20 participants	89.7
RespEar [16, 18]	Breathing rate estimation dataset encompassing 6 activities with 18 participants	632.3
In-ear cardiac sounds [198]	In-ear heart sounds segmentation dataset with different noise levels for 14 participants	89.5
EarMeter [199]	Tidal volume estimation dataset with resting and cooldown (elevated heart and breathing rate) activities for 20 participants	353.2
Total		1164.7

This enables the model to learn from local structures in the spectrogram while still using masking to promote deeper feature learning. To address the dataset imbalance shown in Figure 5.4.2, we augment the dataset by duplicating and discarding samples to retain the median number of samples in the dataset per bin. This prevents the model from being biased towards more frequently occurring labels. We train the model with a learning rate of 1×10^{-4} and a batch size of 16 for 40 epochs.

We use a custom loss function which combines the MSE loss and a weighted MSE loss. The MSE loss, defined in Equation (5.1), penalises errors across all target values equally, which helps the model to perform well on centrally located values. However, in this system, it is equally as important to accurately predict stroke volumes at the ranges of the distribution. To emphasise the importance of these points, we designed a weighted MSE loss which emphasises prediction errors towards the upper and lower range of the collected dataset. The loss is defined according to Equation (5.2) where y_i are the ground truth values and $\hat{y}_i$ are the model predictions.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.1}$$

$$\mathcal{L}_{weighted} = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2 \tag{5.2}$$

In the weighted MSE loss, w_i is the weight assigned to each prediction, as defined in

Equation (5.3) where μ is the midpoint of the dataset. We normalise the weights by dividing them by the mean squared deviation of the lower or upper range, depending on whether y_i is less than or greater than μ. This normalisation ensures that the contributions from both the higher and lower target ranges remain balanced and prevent the weighted loss from disproportionately dominating the optimisation process.

$$w_i = \begin{cases} \frac{|y_i - \mu|^2}{mean(|y_j - \mu|^2 \ for \ y_j \leq \mu)}, & \text{if } y_i \leq \mu \\ \frac{|y_i - \mu|^2}{mean(|y_j - \mu|^2 \ for \ y_j > \mu)}, & if \ y_i > \mu \end{cases} \tag{5.3}$$

The overall loss function, defined in Equation (5.4), is a combination of the MSE loss and the weighted MSE loss, where α is set to 0.5 to balance the contributions of each loss component.

$$\mathcal{L}_{custom} = \alpha \mathcal{L}_{weighted} + (1 - \alpha)\mathcal{L}_{\text{MSE}} \tag{5.4}$$

We fine-tune the model to predict the stroke volume per window and ultimately average together the predictions to compute the average stroke volume for a user. To evaluate model performance, we use leave-one-subject-out cross-validation. In this approach, the model is trained using data from all but one subject and tested on the data from the remaining subject. This method assesses the model's ability to predict average stroke volume for a subject whose data it has not encountered before, mimicking the ideal use case for such a system.

5.4 Implementation

5.4.1 Data collection

A total of 23 healthy volunteers (detailed in Section 5.4.2) were recruited for this study, which aimed to evaluate the feasibility of determining stroke volume using in-ear audio. The study was approved by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge, where data collection was conducted. Written informed consent was obtained from each participant prior to the start of the data collection protocol.

The study was conducted by trained medical professionals. The data collection process,

illustrated in Figure 5.3.1(a), began with participants resting on a bed for 10 minutes. After this rest period, they performed isometric exercises, specifically squeezing a stress ball with quick, forceful contractions for 30 seconds, to elevate heart rate and alter stroke volume. Following the exercises, participants rested for an additional 5 minutes. This protocol was designed to capture stroke volume measurements in the resting, baseline state, and the exercise recovery state to evaluate the accuracy of our method under different conditions. In this analysis, we analyse the data in the resting state and the recovery state.

During the data collection, subjects wore custom earbuds (Figure 3.4.1(a)) in both ears to collect in-ear audio signals, fitted with eartips of varying sizes to ensure proper occlusion. Ground truth stroke volume measurements were obtained using the Finapres NOVA continuous hemodynamic monitoring system [107], which estimates beat-to-beat stroke volume through a PPG finger cuff.

To set up the Finapres, a medical-grade three-lead ECG electrode was attached to each participant's chest, and a blood pressure cuff was placed on their non-dominant arm for calibration. The system was calibrated using stroke volume measurements obtained from an echocardiogram for each participant (Figure 2.5.2). Calibration stroke volume was measured with the GE Vivid IQ ultrasound system, calculated as the product of the left ventricular outflow tract (LVOT) cross-sectional area and the Doppler flow envelope recorded at the LVOT [105]. The average stroke volume was derived from three cardiac cycles. Additionally, the Finapres was calibrated using resting blood pressure measurements from the arm, taken while participants were lying down.

In-ear audio data was collected with a sampling rate of 44 kHz, while Finapres data was recorded on a beat-by-beat basis. The data streams were synchronised using timestamps recorded on both devices.

5.4.2 Study population

Table 5.4.1 provides the demographics for the study population of 23 participants, consisting of 13 men and 10 women across a range of ages, weights and heights. Specifically, participants' ages ranged from 21 to 58 years, and their body mass indexes (BMI) spanned from 19 to 42, covering the range of normal to obese. This variation highlights the high level of demographic diversity within the cohort.

In addition to demographic diversity, participants with varying fitness levels were recruited

Table 5.4.1: Participant Demographics

Demographic	Mean	Min	Max	SD
Age	32	21	58	11
Weight (kg)	76	55	113	15
Height (cm)	176	158	194	9
Body mass index	25	19	42	5
Heart rate (BPM)	69	46	108	9
Stroke volume (ml)	79	45	118	15
Cardiac output (l)	5.41	2.97	11.06	1.09

to ensure a broad distribution of stroke volumes. The group included individuals with diverse exercise habits, from those who rarely exercise to those who work out daily, as illustrated in Figure 5.4.1(a). The relationship between average stroke volume and exercise frequency for the participants is summarised in Figure 5.4.1(b). Generally, the median stroke volume increases with increasing exercise frequency, though there is a greater variation in stroke volumes among those exercising 3-4 times per week compared to other frequencies. This variation may be due to the different types of exercise, as endurance exercises can affect cardiac response differently than high-intensity exercises.

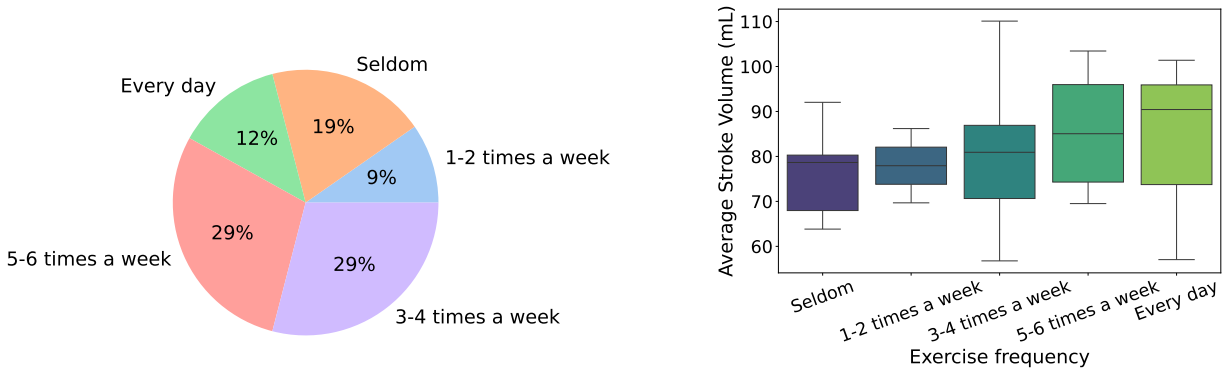
Figure 5.4.2 provides the distribution of beat-by-beat stroke volume data collected in the study. A slight right shift is noticeable in the post-exercise data compared to pre-exercise, indicating a higher number of individual heartbeats with larger stroke volumes. On average, the mean stroke volume before exercise is 77 ± 13 mL, while after exercise, it increased slightly to 79 ± 14 mL. Comparing the average stroke volume per user before and after exercise, using a Wilcoxon signed-rank test, the difference is statistically significant with $p < 0.05$.

5.5 Evaluation

In this section, we present the results of stroke volume estimation using in-ear audio.

5.5.1 Metrics

We evaluate system performance using the MAE between the ground truth stroke volume and the calculated stroke volume, and the MAPE between the two. We also assess performance using a Bland-Altman plot, as described in Chapter 3. In addition, we evaluate



(a) (b)

Figure 5.4.1: Overview of the impact of exercise on the participants' average stroke volume. (a) Breakdown of the exercise frequency of the surveyed participants. (b) Comparison of average stroke volume per exercise frequency.

the correlation between the two measurements using Pearson's correlation analysis, and the level of agreement using the R^2 score (the coefficient of determination). Finally, we assess whether the errors are clinically acceptable by computing the percentage error (PE) which is defined as 1.96 times the standard deviation of the bias from the Bland-Altman plot, divided by the average ground truth stroke volume [104, 187]. For a stroke volume measurement device to be accurate, it must have a PE of less than 30% [13, 104, 187].

5.5.2 Overall stroke volume prediction

Overall, our system can predict stroke volume from in-ear audio signals for unseen subjects with a MAE of 5.24 ml, and a MAPE of 6.83%. Figure 5.5.1 provides the scatter plot of the stroke volume predictions per subject including the identity line (the ideal line where predictions are equal to the ground truth). This plot examines the average stroke volume per subject over the entire experimental protocol. The calculated stroke volume is correlated with the ground truth stroke volume with a Pearson's correlation coefficient (r) of 0.94 ($p<0.001$), indicating a very high agreement between the predictions and the ground truth. The R^2 score of 0.88 indicates that 88% of the variance in the ground truth stroke volume is explained by the predictions, demonstrating a close alignment between the predictions and the true values. From Figure 5.5.1, it is evident that, generally, stroke volume estimations in the centre of the range are highly accurate. Towards the ends of the range, there is more variability with a sight tendency to underestimate stroke volume. Nonetheless, the system predicts average stroke volume, even for new users, with excellent agreement with

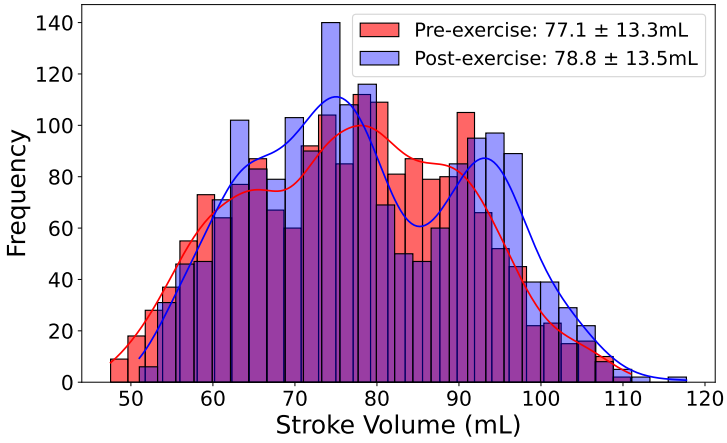


Figure 5.4.2: Distribution of beat-by-beat stroke volumes in the collected data before and after exercise.

the ground truth as indicated by both the low error and the high correlation coefficient.

The overall PE is 11.05%, which is well below the 30% PE deemed acceptable for a new stroke volume measurement device by Critchley and Critchley [200], highlighting the potential of our solution. To further contextualise this result, we compare our performance with the reported results of other wearable devices used in literature in Table 5.5.1. It is clear that our system outperforms all prior works, with significantly lower PE, while additionally using only commodity sensors that are readily available in commercial devices.

Table 5.5.1: Performance comparison between our system and wearable-based systems reported in the literature.

Author	**Modality**	**Percentage Error (%)**
Ours	**Earable**	**11.05**
Ganti et al. [13]	Chest patch with a single-lead electrocardiogram and a triaxial accelerometer	28
Dvir et al. [14]	Chest patch with PPG sensor	25.6
Yadzi et al. [104]	Custom sensor embedded scale	36.73
Wang et al. [15]	Finger-based PPG	16.2

While the current error rate is acceptable and outperforms the literature, it also highlights a limitation of our work. Despite the dataset covering a wide range of stroke volumes, the overall standard deviation is only 15 ml. For comparison, a previous study involving 1450 individuals reported a standard deviation of 18 ml [201], indicating that our dataset

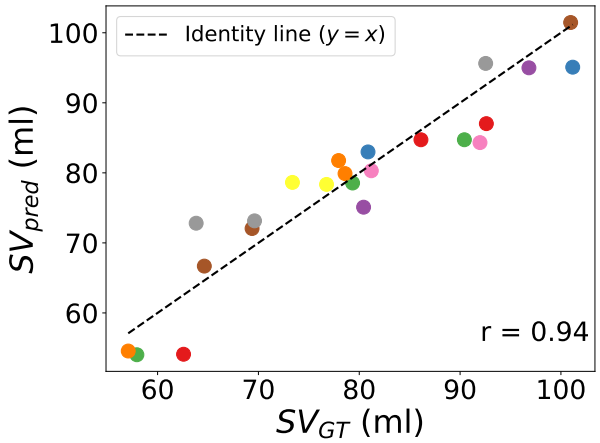


Figure 5.5.1: Scatter plot comparing the overall calculated stroke volume (SV_{pred}) and the ground truth stroke volume (SV_{GT}) per user. The Pearson correlation coefficient r quantifies the quality of correlation.

reflects similar variability to that seen in a larger population. However, from a processing standpoint, a narrower standard deviation means less variability in the dataset, which may limit the model's ability to generalise to stroke volumes further from the mean. This challenge is compounded by the diversity in the in-ear audio features, making it more difficult to accurately predict stroke volumes at the extremes of the range. This trend is evident in Figure 5.5.1, where the smallest estimation errors are found for users with stroke volumes near the dataset average, while errors tend to increase slightly towards the lower and upper ends of the range. Nonetheless, the results demonstrate that our system, utilising self-supervised learning, is robust and effective overall.

To further improve the system's generalisability, future work should focus on expanding the dataset to include a wider distribution of stroke volumes, ages, ethnic groups, and fitness levels, to enhance the model's ability to generalise and minimise the risk of bias. Future work should also focus on collecting data under exercise to fully assess the ability of the system to predict altered stroke volumes that are different from the baseline. Importantly, this expanded dataset will be integrated into our current framework without the need for changes to the already highly effective modelling technique.

5.5.3 Sensitivity to change in exercise condition

Our experimental protocol involves recording data in a baseline resting state (pre-exercise), and then in a post-exercise state to assess the body's response to exercise. The exercise

condition, which occurs between these two assessments, involves isometric exercise designed to produce cardiovascular stress which alters stroke volume. Clinically, assessing stroke volume before and after exercise is crucial for evaluating a patient's cardiovascular health and response to exercise. Therefore, it is important to determine whether the system can accurately predict both the changes in stroke volume due to exercise, and the resting, stable state, where significant variations in stroke volume are unlikely.

Figure 5.5.2 provides a scatter plot of the stroke volume predictions per user per condition. The overall correlation between the ground truth stroke volume and calculated stroke volume of the pre-exercise and post-exercise conditions is 0.93 ($p<0.001$) with an R^2 score of 0.87, similar to that of the overall predictions. From the Bland-Altman analysis in Figure 5.5.3, we achieve a bias of -0.69 mL and limits of agreement of -10.02 mL to 8.63 mL in the combined condition. The very low bias and narrow limits of agreement (which contain most of the data points) again indicate a strong agreement between the predictions and the measurements. However, the negative bias shows a slight tendency to underestimate, as was observed in Figure 5.5.1. There is no evidence of a difference in pattern between the pre and post-exercise conditions which suggests that the model's performance is consistent across the two conditions. From the plot, we compute the overall PE as 11.97%, again lower than 30%.

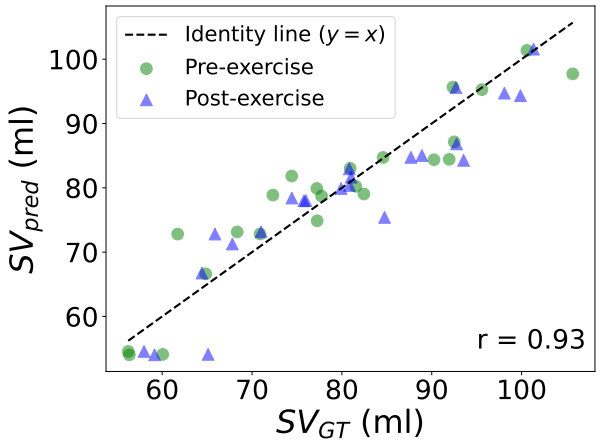


Figure 5.5.2: Relationship between the calculated stroke volume and the ground truth stroke volume for the two exercise conditions: before exercise (green circles) and after exercise (blue triangles).

Separating the pre and post-exercise conditions, a correlation of 0.94 ($p<0.001$) before exercise, and 0.93 ($p<0.001$) after exercise is achieved. Likewise, before exercise, there

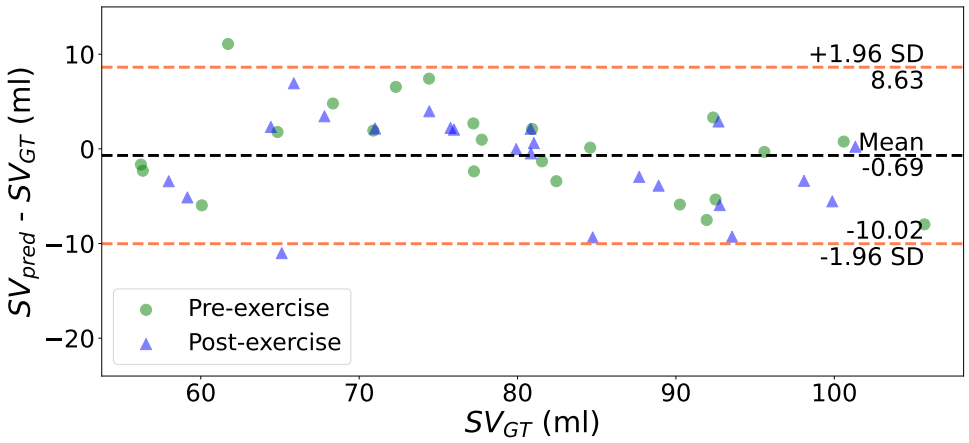


Figure 5.5.3: Bland-Altman plot showing the agreement between the calculated and ground truth stroke volumes. Markers correspond to the two exercise conditions: before exercise (green circles) and after exercise (blue triangles).

is a bias of -0.02mL and limits of agreement of -9.35 mL to 9.31 mL, while after exercise, the bias is -1.37mL and limits of agreement are -10.50 mL to 7.77 mL. This shows that the system underestimates more in the post-exercise condition where stroke volumes are elevated. However, this tendency is minor and can be overcome in future by collecting more data with elevated stroke volumes by including longer exercise durations. From the bias and limits of agreement, our system achieves a PE of 12.06% and 11.65% for the pre and post-exercise conditions respectively. The system achieves a MAE (MAPE) of 5.27 mL (6.81%) and 5.21 mL (6.85%) before and after exercise respectively. The similarity in errors demonstrates the system's ability to accurately predict stroke volume under both altered conditions post-exercise and in resting conditions, proving the system's generalisability in predicting a range of stroke volume responses.

Moreover, the estimation error per participant for the pre and post-exercise conditions is provided in Figure 5.5.4. Overall, only two participants (2 and 21) have a MAPE of over 10%, highlighting the generalisability of our system. If we separate the two conditions, most users have a MAPE of less than 10%, showing that predictions are accurate for both conditions. However, for four participants, errors of over 10% are seen for one or both conditions. For participant 2, the high MAPE is due to a slight under-prediction of the post-exercise condition, which manifests as a large error due to the low ground truth value. Likewise, for participant 9, the pre-exercise stroke volume is accurately predicted but underestimated after exercise. In participant 13, the pre-exercise stroke volume is

over-estimated. Participant 21 has the highest overall error of all participants with errors of above 10% in both the pre and post-exercise conditions. This is due to the participant having a low ground truth stroke volume which is over-estimated leading to a larger percentage error. This again highlights the large inter-participant variability in in-ear audio signals, which we expect will be mitigated by collecting data from more participants. Regardless, the performance is excellent for the majority of users.

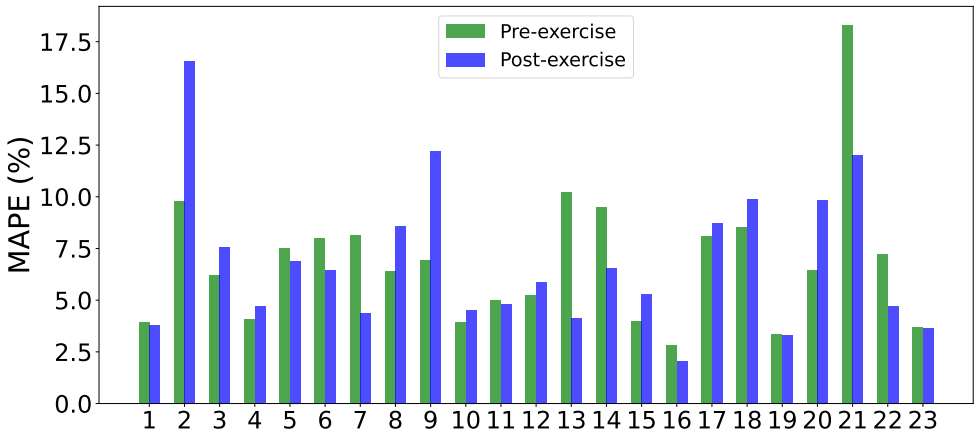


Figure 5.5.4: Per-participant estimation MAPE for the two conditions.

5.5.4 Stroke volume estimation using demographics and audio features

Height, weight, body surface area and body mass index have been proven to be moderately correlated with stroke volume [202, 203]. To validate the strength of our technique, we ran a study to predict the average stroke volume per user using only their demographics (age, weight, height, and sex). Using a support vector regressor with demographics as features, stroke volume can be estimated with a MAE of 11.24 mL (15% error), with a correlation of 0.40 (p=0.44). Thus, although there is a moderate correlation between the predictions and the labels, it is not statistically significant, indicating no evidence of a linear relationship between the two. Consequently, demographics alone cannot be used to estimate stroke volume.

Heart rate also has been shown to have a high feature importance for stroke volume predictions [13]. As such, we use average heart rate to predict average stroke volume, achieving a MAE of 10.72 mL with an error of 14% (r=0.2, not statistically significant),

showing that heart rate is slightly more predictive of stroke volume than demographics, yet both achieve poor performance.

We further attempted to predict stroke volume within 10s windows using the features extracted in Figure 5.2.2 combined with the demographics. We compute the average predicted stroke volume and compare it to the average ground truth stroke volume. Using the best-performing regressor (the support vector regressor with radial basis function kernel), we achieve similar results with a MAE of 11.82 mL (15% error). Thus, this study confirms that feature extraction alone cannot achieve generalisable results on unseen users, and confirms the need for a more sophisticated system, as implemented in this work.

5.5.5 Stroke volume estimation without using self-supervised pre-training

In this section, we evaluate the importance of our pre-training technique. Using the weights obtained by training AudioMAE on Audioset (as described in Section 5.3.3), we train the model for each user as in Section 5.3.4 and evaluate performance without self-supervised pre-training. Without pre-training, the model achieves a MAE of 9.72 mL with a MAPE of 13.01%, a Pearson correlation of 0.77 and a PE of 29.18%. In contrast, by incorporating self-supervised pre-training, we observe a significant performance improvement of 48%, demonstrating that the pre-training technique has effectively enhanced the model's generalisability.

5.6 Conclusion

This chapter presented CardiAural, the first earable-based system for stroke volume estimation. CardiAural uses a generative self-supervised learning framework with a masked autoencoder architecture and transfer learning to predict average stroke volume before and after isometric exercise. We collected a dataset from 23 participants using our earable prototype with a clinically validated stroke volume estimation device as ground truth. Our results show that we achieve percentage errors within the acceptable range for a new stroke volume measurement device with very good agreement between stroke volume predictions and ground truth labels. Our excellent performance demonstrates, for the first time, the feasibility of stroke volume monitoring using commodity in-ear microphones commonly found in earables.

Chapter 6

Conclusions and Discussion

In this thesis, we have presented a body of work exploring the use of in-ear audio from earables for physiological monitoring. Through this work, we address fundamental challenges in earable research for physiological monitoring: (i) overcoming the low signal-to-noise ratio of biosignals and their susceptibility to motion artefacts to derive meaningful measurements from low-quality data, (ii) developing effective and efficient processing pipelines with low latency and power consumption, and (iii) creating models that can generalise to different signal properties and users despite limited sized datasets. Ultimately, this thesis demonstrates how the unique properties of in-ear audio, combined with novel processing pipelines, can enable the monitoring of various critical physiological parameters such as heart rate, respiratory rate, and stroke volume. In this concluding chapter, we briefly summarise our key contributions to answer the research questions presented in Chapter 1. We then discuss the limitations of this research and propose future research directions beyond this thesis.

6.1 Summary of Contributions

6.1.1 Motion-resilient heart rate monitoring

In Chapter 3, we explored the use of in-ear audio for heart rate monitoring in both sedentary and active conditions. We performed an analysis of the interference caused by full-body motion and found that it creates significant motion artefacts that overlap in frequency with both heart sounds and heart rate. To address this challenge, we presented a deep

learning-based denoising pipeline for heart rate estimation, using transfer learning with a U-Net model to denoise and enhance heart sounds in in-ear audio using ECG as a reference. After a comprehensive data collection and analysis, we demonstrated that our system could predict heart rate with errors of less than 10% while users were sedentary, walking and running. We proved that our system achieves accurate results in both indoor and outdoor environments, while listening to music, and longitudinally. These results indicate that commodity earbuds equipped with in-ear microphones can be used to accurately estimate heart rate across a range of real-world settings, highlighting the potential of earable devices for heart rate monitoring.

6.1.2 Robust respiratory rate monitoring

In Chapter 4, we investigated the feasibility of using in-ear audio to measure respiratory rate, even in the absence of audible breathing sounds, under both sedentary and active conditions. We performed an analysis of the common sensors on earables for respiratory rate sensing and showed that none of the sensors can directly detect breathing sounds or breathing motions under all conditions. However, we showed that the in-ear microphone is the best suited for respiratory rate sensing due to its ability to capture both heart sounds and footstep sounds. We proposed a processing pipeline that leverages heart and footstep sounds to indirectly estimate respiratory rate using physiological couplings, specifically Respiratory Sinus Arrhythmia and Locomotor Respiratory Coupling. Our system consists of two parallel signal-processing pipelines depending on the user's activity level: while sedentary, respiratory rate is extracted from heart rate variability signals, and when active, respiratory rate is extracted from audio features related to respiration and stride frequency. We showed that this approach can reliably and accurately estimate respiratory rate across various daily life activities and conditions, outperforming existing methods for respiratory rate monitoring on wearable devices. Thus, we demonstrated the feasibility of using a single sensor on a commodity wearable device for robust respiratory rate monitoring.

6.1.3 Stroke volume monitoring

In Chapter 5, we examined the possibility of measuring stroke volume using in-ear audio, thus extending our work to a vital sign typically only measured in clinical settings. We investigated the relationship between common audio features and demographic factors with stroke volume, finding weak correlations that were not generalisable to diverse signal

properties across participants. To overcome this, we proposed a pipeline that leverages generative self-supervised learning using a masked autoencoder with transfer learning to predict average stroke volume before and after isometric exercise. Our approach involved pre-training the model on a dataset of unlabelled in-ear audio, allowing it to learn the characteristic features of the audio. We then fine-tuned the model on a specialised, labelled dataset to predict stroke volume. Our results showed that stroke volume can be accurately estimated from in-ear audio under both pre and post-exercise conditions, paving the way to unlocking widespread monitoring of cardiovascular health and fitness.

6.2 Limitations and Future Research Directions

The work presented in this thesis has promising implications for the mobile computing community and beyond. Researchers can build upon our methods and ideas to further develop scalable, generalisable and robust sensing systems. From a commercial perspective, engineers could embed our pipelines and models into existing noise-cancelling earbuds and hearing aids, creating the next generation of smart wearables with embedded sensing capabilities from the ear. This would allow consumers to better track their fitness, health and wellbeing. Additionally, medical practitioners could use our systems to longitudinally monitor patients outside of clinical settings, providing health data from settings which are more representative of their everyday lives. The ability of individuals and doctors to monitor health continuously could not only impact personal lifestyles but also influence and inform public health policies. However, all research has limitations and it is essential to uncover these so that future work can be done to improve upon them. In the remainder of this section, we will examine the limitations of the presented work and potential solutions for future development.

6.2.1 Dataset size

One of the main limitations of the work discussed is the small size of the collected datasets. While our systems performed well on these datasets, larger datasets are needed to verify their effectiveness at scale and across more diverse demographics. Moreover, larger datasets are expected to produce better results, as they more accurately represent the underlying data distribution, thereby improving the ability of the models and systems to generalise. However, collecting large datasets presents a significant challenge in earable research. Firstly, there is a lack of commercial earable devices that provide access to APIs for onboard sensors,

specifically inward-facing microphones. As a result, we had to develop a custom earable prototype for our data collections. This meant that devices could not be used in the wild or without investigators present, thus limiting the amount of data that could be collected. Additionally, custom hardware is fragile and prone to breakage, resulting in data corruption and sometimes poor data quality. Over the course of the three studies in this thesis, many participants' data had to be discarded due to these issues. Secondly, since data collection sessions had to be conducted under controlled conditions, the sessions were time-consuming and required participants to travel to the investigators. This limits the number of participants that can be recruited and as a result, the dataset size. A further limitation of the presented work is the limited set of activities used to evaluate the algorithms. Although effort was taken to collect data from a large number of activities, these activities are not exhaustive, and thus the results presented in Chapters 3 and 4 are not indicative of true in-the-wild performance.

For future work, efforts should be taken to collect data from a wider number of participants over longer periods. These collections should also take place in the wild, where a participant is given a ground truth device and an earable which they wear as they go about their daily activities. This will enable better generalisation of the systems to varying activities, degrees of motion artefacts and ambient conditions, and also minimise the risk of model bias. It is also critical to expand upon the demographics represented in the datasets to better represent a range of ages, genders and ethnic groups. Further, for all work in this thesis, but particularly for the work presented in Chapter 5, it would be valuable to collect data from participants with health conditions, particularly those related to the cardiovascular system.

6.2.2 Earable fit and seal quality

Throughout the works explored in this dissertation, earable fit is a recurring limitation. Due to our reliance on the occlusion effect, achieving a good seal is essential for capturing high-quality data. In cases where a proper seal was not obtained, the data from some users had to be excluded because the signal was of poor quality and thus unusable. To address this, future systems should implement an automated test of seal quality that informs users to try a different-sized ear tip or adjust the earbud for a better fit when poor quality signals are encountered. This can be achieved using the technique in [204], which measures the reduction in amplitude of predefined frequencies to determine whether the earbud is

properly sealed against external noise. Alternatively, techniques for amplifying weak signals in the presence of partial occlusion should be explored.

6.2.3 Personalisation

An additional aspect left unexplored is whether personalisation can help improve the estimation of physiological parameters. In this scheme, models could be fine-tuned using data from individual users to improve estimation performance. Due to the variations in in-ear heart sounds across the population, we expect that training the model using some of the user's own data would improve prediction accuracy, particularly for users whose data significantly differs from that of the average population. However, this approach introduces an additional challenge of how to collect labels for this data. One potential solution is the use of weak labels, such as heart rate data from another device (*e.g.,* a smartwatch) to fine-tune the model for denoising and heart rate estimation. For stroke volume estimation, the label could be obtained from an echocardiogram at an appointment with a physician.

6.2.4 Sensor fusion

In this thesis, we focus exclusively on the in-ear microphone for developing sensing pipelines. However, commodity earbuds also contain IMUs and outer-ear microphones which could be used in conjunction with the in-ear microphone to enhance sensing capabilities. For example, research could be done into denoising in-ear heart signals under motion using the IMU which accurately captures movement. Additionally, the out-ear microphone could provide context about a user's environment, while the IMU could determine the user's activity level to select the appropriate processing pipeline: for example, HRV while stationary, or motion-resilient heart rate during activity.

6.2.5 Looking forward on earables for health and wellbeing

Ultimately, in this thesis, we aimed to develop techniques for monitoring various aspects of human health and wellbeing using commodity earbuds. While we have proven, for the first time, that heart rate, respiratory rate, and stroke volume can be accurately monitored using in-ear microphones, more work is needed to transfer these systems from controlled settings to real-world environments. Due to the limitations of existing devices, we were unable to deploy our systems onto commercial devices to assess their real-world feasibility, thus limiting the practical impact of our research. Future efforts should focus on real-world

deployment to achieve the overall goal of monitoring human physiology with earables in the wild.

This work also opens up the possibility of monitoring many other aspects of human health and wellbeing using earables. Accurate heart rate monitoring leads to the question of whether HRV can be monitored under motion. The ability to detect heart sounds through the ear suggests that earables could potentially be used to detect heart murmurs or other cardiac abnormalities. Given the close link between the ears and the respiratory system, future research could also explore the possibility of detecting respiratory illnesses, such as tuberculosis, using earables.

Finally, earbuds offer unique opportunities for sensing and user interaction since they can simultaneously sense and provide audio-based interventions to the user. Future efforts should focus on developing end-to-end systems that not only sense physiological phenomena but also provide feedback to inform behavioural change to help users improve their health and fitness.

It is often said that the eyes are the window to the soul. I believe that the ears are the windows to the body, and I look forward to seeing what new aspects of health and physiology can be unlocked using earables in the years to come.

Bibliography

[1] V. Goverdovsky, W. von Rosenberg, T. Nakamura, D. Looney, D. J. Sharp, C. Papavassiliou, M. J. Morrell, and D. P. Mandic, "Hearables: Multimodal physiological in-ear sensing," *Scientific Reports*, vol. 7, no. 1, p. 6948, Dec. 2017.

[2] F. Kawsar, C. Min, A. Mathur, and A. Montanari, "Earables for Personal-Scale Behavior Analytics," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 83–89, Jul. 2018.

[3] F. Laricchia, "Global wearable shipments by category 2028," https://www.statista.com/statistics/1265326/wearables-worldwide-shipments-quarterly-by-product-category/, [Accessed 23-09-24].

[4] A. Ferlini, A. Montanari, C. Min, H. Li, U. Sassi, and F. Kawsar, "In-Ear PPG for Vital Signs," *IEEE Pervasive Computing*, vol. 21, no. 1, pp. 65–74, Jan. 2022.

[5] X. Fan, D. Pearl, R. Howard, L. Shangguan, and T. Thormundsson, "APG: Audioplethysmography for Cardiac Monitoring in Hearables," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. Madrid Spain: ACM, Oct. 2023, pp. 1–15.

[6] "AirPods Pro (2nd generation) - Technical Specifications," https://www.apple.com/uk/airpods-pro/specs/, 2023, [Accessed 23-09-24].

[7] "WF-1000XM5 Wireless Noise Cancelling Headphones," https://www.sony.co.uk/headphones/products/wf-1000xm5/features1, [Accessed 23-09-24].

[8] M. M. Rahman, X. Xu, V. Nathan, T. Ahmed, M. Y. Ahmed, D. McCaffrey, J. Kuang, T. Cowell, J. Moore, W. B. Mendes, and J. A. Gao, "Detecting Physiological Responses Using Multimodal Earbud Sensors," in *2022 44th Annual International Conference of*

the IEEE Engineering in Medicine & Biology Society (EMBC). Glasgow, Scotland, United Kingdom: IEEE, Jul. 2022, pp. 01–05.

[9] W. von Rosenberg, T. Chanwimalueang, V. Goverdovsky, N. S. Peters, C. Papavassiliou, and D. P. Mandic, "Hearables: Feasibility of recording cardiac rhythms from head and in-ear locations," *Royal Society Open Science*, vol. 4, no. 11, p. 171214, Nov. 2017.

[10] A. Martin and J. Voix, "In-Ear Audio Wearable: Measurement of Heart and Breathing Rates for Health and Safety Monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1256–1263, 2018.

[11] Y. Nam, B. A. Reyes, and K. H. Chon, "Estimation of Respiratory Rates Using the Built-in Microphone of a Smartphone or Headset," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1493–1501, Nov. 2016.

[12] Y. Ren, C. Wang, J. Yang, and Y. Chen, "Fine-grained sleep monitoring: Hearing your breathing with smartphones," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 1194–1202.

[13] V. G. Ganti, A. H. Gazi, S. An, A. V. Srivatsa, B. N. Nevius, C. J. Nichols, A. M. Carek, M. Fares, M. Abdulkarim, T. Hussain, F. G. Greil, M. Etemadi, O. T. Inan, and A. Tandon, "Wearable Seismocardiography-Based Assessment of Stroke Volume in Congenital Heart Disease," *Journal of the American Heart Association*, vol. 11, no. 18, p. e026067, Sep. 2022.

[14] A. Dvir, N. Goldstein, A. Rapoport, R. G. Balmor, D. Nachman, R. Merin, M. Fons, A. Ben Ishay, and A. Eisenkraft, "Comparing Cardiac Output Measurements Using a Wearable, Wireless, Noninvasive Photoplethysmography-Based Device to Pulse Contour Cardiac Output in the General ICU: A Brief Report," *Critical Care Explorations*, vol. 4, no. 2, p. e0624, Feb. 2022.

[15] L. Wang, E. Pickwell-MacPherson, Y. Liang, and Y. Zhang, "Noninvasive cardiac output estimation using a novel photoplethysmogram index," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Minneapolis, MN: IEEE, Sep. 2009, pp. 1746–1749.

[16] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, Y. Liu, and C. Mascolo, "An evaluation of heart rate monitoring with in-ear microphones under motion," *Pervasive and Mobile*

Computing, p. 101913, 2024.

[17] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, and C. Mascolo, "hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones," in *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mar. 2023, pp. 200–209.

[18] Y. Liu, K.-J. Butkow, J. Stuchbury-Wass, A. Pullin, D. Ma, and C. Mascolo, "RespEar: Earable-based robust respiratory rate monitoring," 2024.

[19] K.-J. Butkow, N. Jalaludeen, Y. Liu, J. Stuchbury-Wass, Q. Yang, M. Ciliberto, D. Ma, J. Cheriyan, and C. Mascolo, "Measuring cardiac stroke volume through in-ear audio sensing," 2024.

[20] K.-J. Butkow, A. Ferlini, F. Kawsar, C. Mascolo, and A. Montanari, "EarTune: Exploring the physiology of music listening," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 644–649.

[21] Q. Yang, Y. Liu, J. Stuchbury-Wass, K.-J. Butkow, D. Ma, and C. Mascolo, "BrushBuds: Toothbrushing tracking using earphone imus," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 655–660.

[22] T. Ketmalasiri, Y. Y. Wu, K.-J. Butkow, C. Mascolo, and Y. Liu, "IMChew: Chewing analysis using earphone inertial measurement units," in *Proceedings of the Workshop on Body-Centric Computing Systems*, ser. BodySys '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 29–34.

[23] J. Stuchbury-Wass, E. Bondareva, K.-J. Butkow, S. Šćepanović, Z. Radivojevic, and C. Mascolo, "Heart rate extraction from abdominal audio signals," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[24] H. J. Davies, I. Williams, N. S. Peters, and D. P. Mandic, "In-Ear SpO2: A Tool for Wearable, Unobtrusive Monitoring of Core Blood Oxygen Saturation," *Sensors*, vol. 20, no. 17, p. 4879, Aug. 2020.

[25] ——, "In-ear measurement of blood oxygen saturation: An ambulatory tool needed to detect the delayed life-threatening hypoxaemia in COVID-19," *arXiv:2006.04231 [eess]*, Jun. 2020.

[26] K. Matsumoto, Y. Temiz, H. Taghavi, E. L. Cornelius, H. Mori, and B. Michel, "An earbud-type wearable (A hearable) with vital parameter sensors for early detection and prevention of heat-stroke," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, Jul. 2019, pp. 7049–7055.

[27] N. Bui, N. Pham, J. J. Barnitz, Z. Zou, P. Nguyen, H. Truong, T. Kim, N. Farrow, A. Nguyen, J. Xiao, R. Deterding, T. Dinh, and T. Vu, "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*. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 1–17.

[28] S. F. LeBoeuf, M. E. Aumer, W. E. Kraus, J. L. Johnson, and B. Duscha, "Earbud-based sensor for the assessment of energy expenditure, heart rate, and vo2max," *Medicine and science in sports and exercise*, vol. 46, no. 5, pp. 1046–1052, May 2014.

[29] E. Bondareva, E. R. Hauksdóttir, and C. 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*, ser. UbiComp '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 146–151.

[30] K. Taniguchi, H. Kondo, M. Kurosawa, and A. Nishikawa, "Earable TEMPO: A novel, hands-free input device that uses the movement of the tongue measured with a wearable ear sensor," *Sensors*, vol. 18, no. 3, p. 733, Mar. 2018.

[31] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *International Conference on Ubiquitous Computing*. Springer, 2005, pp. 56–72.

[32] R. Lotfi, G. Tzanetakis, R. Eskicioglu, and P. Irani, "A comparison between audio and IMU data to detect chewing events based on an earable device," in *Proceedings of the 11th Augmented Human International Conference*. Winnipeg Manitoba Canada: ACM, May 2020, pp. 1–8.

[33] S. Päßler, M. Wolff, and W.-J. Fischer, "Food intake monitoring: An acoustical approach to automated food intake activity detection and classification of consumed food," *Physiological Measurement*, vol. 33, no. 6, p. 1073, May 2012.

[34] E. Nemati, S. Zhang, T. Ahmed, M. M. Rahman, J. Kuang, and A. Gao, "CoughBuddy: Multi-modal cough event detection using earbuds platform," in *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2021, pp. 1–4.

[35] Z. Zhao, F. Li, Y. Xie, Y. Wu, and Y. Wang, "BSMonitor: Noise-resistant bowel sound monitoring via earphones," *IEEE Transactions on Mobile Computing*, vol. 23, no. 4, pp. 3213–3227, 2024.

[36] Y. Jin, Y. Gao, X. Guo, J. Wen, Z. Li, and Z. Jin, "EarHealth: An earphone-based acoustic otoscope for detection of multiple ear diseases in daily life," in *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, ser. MobiSys '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 397–408.

[37] M. A. Stone, A. M. Paul, P. Axon, and B. C. Moore, "A technique for estimating the occlusion effect for frequencies below 125 Hz," *Ear and Hearing*, vol. 35, no. 1, pp. 49–55, Jan. 2014.

[38] J. Tonndorf, "A new concept of bone conduction," *Archives of Otolaryngology*, vol. 87, no. 6, pp. 595–600, 1968.

[39] S. Stenfelt and S. Reinfeldt, "A model of the occlusion effect with bone-conducted stimulation," *International Journal of Audiology*, vol. 46, no. 10, pp. 595–608, Jan. 2007.

[40] K. Carillo, O. Doutres, and F. 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*, vol. 147, no. 5, pp. 3476–3489, May 2020.

[41] W. Commons, "Human ear anatomy with detailed diagram," 2003, [Accessed 21-08-24]. [Online]. Available: https://commons.wikimedia.org/wiki/File:Ear-anatomy-text-small-en.svg

[42] D. Ma, A. Ferlini, and C. Mascolo, "OESense: Employing occlusion effect for in-Ear human sensing," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 175–187.

[43] J. Prakash, Z. Yang, Y.-L. Wei, H. Hassanieh, and R. R. Choudhury, "EarSense: Earphones as a teeth activity sensor," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. London United Kingdom: ACM, Sep. 2020, pp. 1–13.

[44] Y. Cao, C. Cai, A. Yu, F. Li, and J. Luo, "EarAcE: Empowering Versatile Acoustic Sensing via Earable Active Noise Cancellation Platform," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 2, pp. 1–23, Jun. 2023.

[45] A. Ferlini, D. Ma, R. Harle, and C. Mascolo, "EarGate: Gait-based user identification with in-ear microphones," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 337–349.

[46] X. Sun, J. Xiong, C. Feng, W. Deng, X. Wei, D. Fang, and X. Chen, "Earmonitor: In-ear Motion-resilient Acoustic Sensing Using Commodity Earphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–22, Dec. 2022.

[47] K. Christofferson, X. Chen, Z. Wang, A. Mariakakis, and Y. Wang, "Sleep sound classification using ANC-enabled earbuds," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, 2022, pp. 397–402.

[48] F. Han, P. Yang, Y. Feng, W. Jiang, Y. Zhang, and X.-Y. Li, "EarSleep: In-ear acoustic-based physical and physiological activity recognition for sleep stage detection," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 2, May 2024.

[49] W. Xie, Q. Hu, J. Zhang, and Q. Zhang, "EarSpiro: Earphone-based spirometry for lung function assessment," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 4, Jan. 2023.

[50] S. Michael, K. S. Graham, and G. M. Davis, "Cardiac Autonomic Responses during Exercise and Post-exercise Recovery Using Heart Rate Variability and Systolic Time

Intervals-A Review," *Frontiers in Physiology*, vol. 8, p. 301, 2017.

[51] S. Ismail, U. Akram, and I. Siddiqi, "Heart rate tracking in photoplethysmography signals affected by motion artifacts: A review," p. 5, Dec. 2021.

[52] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *npj Digital Medicine*, vol. 3, no. 1, p. 18, Dec. 2020.

[53] J. W. Navalta, J. Montes, N. G. Bodell, R. W. Salatto, J. W. Manning, and M. DeBeliso, "Concurrent heart rate validity of wearable technology devices during trail running," *Plos one*, vol. 15, no. 8, 2020.

[54] R. Al-Halawani, P. H. Charlton, M. Qassem, and P. A. Kyriacou, "A review of the effect of skin pigmentation on pulse oximeter accuracy," *Physiological Measurement*, vol. 44, no. 5, p. 05TR01, May 2023.

[55] G. Chen, S. A. Imtiaz, E. Aguilar–Pelaez, and E. Rodriguez–Villegas, "Algorithm for heart rate extraction in a novel wearable acoustic sensor," *Healthcare technology letters*, vol. 2, no. 1, pp. 28–33, 2015.

[56] P. Sharma, S. A. Imtiaz, and E. Rodriguez-Villegas, "Acoustic Sensing as a Novel Wearable Approach for Cardiac Monitoring at the Wrist," *Scientific Reports*, vol. 9, no. 1, p. 20079, Dec. 2019.

[57] B. G. Rosa, S. Anastasova, and B. Lo, "Small-form wearable device for long-term monitoring of cardiac sounds on the body surface," in *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2021, pp. 1–4.

[58] A. Ferlini, A. Montanari, C. Min, H. Li, U. Sassi, and F. Kawsar, "In-ear PPG for vital signs," *IEEE Pervasive Computing*, vol. 21, no. 1, pp. 65–74, 2022.

[59] S. Nirjon, R. F. Dickerson, Q. Li, P. Asare, J. A. Stankovic, D. Hong, B. Zhang, X. Jiang, G. Shen, and F. Zhao, "Musicalheart: A hearty way of listening to music," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, 2012, pp. 43–56.

[60] F. R. Gilliam, R. Ciesielski, K. Shahinyan, P. Shakya, J. Cunsolo, J. M. Panchal, B. Król-Józaga, M. Król, O. Kierul, C. Bridges, C. Shen, C. E. Waldman, M. Ring,

T. Szepieniec, A. Barnacka, and S. P. Bhavnani, "In-ear infrasonic hemodynography with a digital health device for cardiovascular monitoring using the human audiome," *npj Digital Medicine*, vol. 5, no. 1, p. 189, Dec. 2022.

[61] X. Fan, L. Shangguan, S. Rupavatharam, Y. Zhang, J. Xiong, Y. Ma, and R. Howard, "HeadFi: Bringing intelligence to all headphones," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 147–159.

[62] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The Importance of Respiratory Rate Monitoring: From Healthcare to Sport and Exercise," *Sensors (Basel, Switzerland)*, vol. 20, no. 21, p. 6396, Nov. 2020.

[63] L. Innocenti, C. Romano, G. Greco, S. Nuccio, A. Bellini, F. Mari, S. Silvestri, E. Schena, M. Sacchetti, C. Massaroni, and A. Nicolò, "Breathing monitoring in soccer: Part I—validity of commercial wearable sensors," *Sensors*, vol. 24, no. 4571, 2024.

[64] N. R. Adão Martins, S. Annaheim, C. M. Spengler, and R. M. Rossi, "Fatigue Monitoring Through Wearables: A State-of-the-Art Review," *Frontiers in Physiology*, vol. 12, p. 790292, Dec. 2021.

[65] "Zephyr bioharness 3.0 chest strap," https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf, 2023.

[66] X. Wang, R. Huang, C. Yang, and S. Mao, "Smartphone sonar-based contact-free respiration rate monitoring," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 2, pp. 1–26, 2021.

[67] Y. Nam, Y. Kong, B. Reyes, N. Reljin, and K. H. Chon, "Monitoring of heart and breathing rates using dual cameras on a smartphone," *PloS one*, vol. 11, no. 3, p. e0151013, 2016.

[68] H. Aly and M. Youssef, "Zephyr: Ubiquitous accurate multi-sensor fusion-based respiratory rate estimation using smartphones," in *IEEE INFOCOM 2016-the 35th Annual IEEE International Conference on Computer Communications*. IEEE, 2016, pp. 1–9.

[69] M. M. Rahman, E. Nemati, V. Nathan, and J. Kuang, "Instantrr: Instantaneous respiratory rate estimation on context-aware mobile devices," in *13th EAI International Conference on Body Area Networks 13*. Springer, 2020, pp. 267–281.

[70] S. Valentine, A. C. Cunningham, B. Klasmer, M. Dabbah, M. Balabanovic, M. Aral, D. Vahdat, and D. Plans, "Smartphone movement sensors for the remote monitoring of respiratory rates: Technical validation," *DIGITAL HEALTH*, vol. 8, p. 205520762210890, Jan. 2022.

[71] J. Hernandez, D. McDuff, and R. Picard, "BioWatch: Estimation of Heart and Breathing Rates from Wrist Motions," in *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. Istanbul, Turkey: ICST, 2015.

[72] X. Sun, L. Qiu, Y. Wu, Y. Tang, and G. Cao, "SleepMonitor: Monitoring Respiratory Rate and Body Position During Sleep Using Smartwatch," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–22, Sep. 2017.

[73] D. Liaqat, M. Abdalla, P. Abed-Esfahani, M. Gabel, T. Son, R. Wu, A. Gershon, F. Rudzicz, and E. D. Lara, "WearBreathing: Real World Respiratory Rate Monitoring Using Smartwatches," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–22, Jun. 2019.

[74] T. Hao, C. Bi, G. Xing, R. Chan, and L. Tu, "MindfulWatch: A smartwatch-based system for real-time respiration monitoring during meditation," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–19, 2017.

[75] "Garmin," 2023.

[76] "Apple watch," 2023.

[77] L. Zhao, F. Zhang, H. Zhang, Y. Liang, A. Zhou, and H. Ma, "Robust Respiratory Rate Monitoring Using Smartwatch Photoplethysmography," *IEEE Internet of Things Journal*, vol. 10, no. 6, pp. 4830–4844, Mar. 2023.

[78] R. Dai, C. Lu, M. Avidan, and T. Kannampallil, "Respwatch: Robust measurement of respiratory rate on smartwatches with photoplethysmography," in *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, 2021, pp. 208–220.

[79] B. Paliakaitė, A. Petrėnas, A. Sološenko, and V. Marozas, "Modeling of artifacts in the wrist photoplethysmogram: Application to the detection of life-threatening

arrhythmias," *Biomedical Signal Processing and Control*, vol. 66, p. 102421, Apr. 2021.

[80] K. B. Kim and H. J. Baek, "Photoplethysmography in wearable devices: A comprehensive review of technological advances, current challenges, and future directions," *Electronics*, vol. 12, no. 2923, 2023.

[81] S. Yue, H. He, H. Wang, H. Rahul, and D. Katabi, "Extracting Multi-Person Respiration from Entangled RF Signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–22, Jul. 2018.

[82] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart Homes that Monitor Breathing and Heart Rate," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul Republic of Korea: ACM, Apr. 2015, pp. 837–846.

[83] J. Liu, Y. Zeng, T. Gu, L. Wang, and D. Zhang, "WiPhone: Smartphone-based Respiration Monitoring Using Ambient Reflected WiFi Signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–19, Mar. 2021.

[84] A. Natarajan, H.-W. Su, C. Heneghan, L. Blunt, C. O'Connor, and L. Niehaus, "Measurement of respiratory rate using wearable devices and applications to COVID-19 detection," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–10, 2021.

[85] P. H. Charlton, D. A. Birrenkott, T. Bonnici, M. A. Pimentel, A. E. Johnson, J. Alastruey, L. Tarassenko, P. J. Watkinson, R. Beale, and D. A. Clifton, "Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review," *IEEE reviews in biomedical engineering*, vol. 11, pp. 2–20, 2017.

[86] W. Karlen, C. J. Brouse, E. Cooke, J. M. Ansermino, and G. A. Dumont, "Respiratory rate estimation using respiratory sinus arrhythmia from photoplethysmography," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1201–1204.

[87] Z. Xu, T. Sakagawa, A. Furui, S. Jomyo, M. Morita, M. Ando, and T. Tsuji, "Toward a robust estimation of respiratory rate using cardiovascular biomarkers: Robustness analysis under pain stimulation," *IEEE Sensors Journal*, vol. 22, no. 10, pp. 9904–9913, 2022.

[88] A. Schäfer and K. W. Kratky, "Estimation of breathing rate from respiratory sinus arrhythmia: Comparison of various methods," *Annals of Biomedical Engineering*, vol. 36, pp. 476–485, 2008.

[89] K. Taniguchi and A. Nishikawa, "Earable POCER: Development of a Point-of-Care Ear Sensor for Respiratory Rate Measurement," *Sensors*, vol. 18, no. 9, p. 3020, Sep. 2018.

[90] J. Romero, A. Ferlini, D. Spathis, T. Dang, K. Farrahi, F. Kawsar, and A. Montanari, "OptiBreathe: An Earable-based PPG System for Continuous Respiration Rate, Breathing Phase, and Tidal Volume Monitoring," in *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. San Diego CA USA: ACM, Feb. 2024, pp. 99–106.

[91] A. Bestbier and P. R. Fourie, "Development of a vital signs monitoring wireless ear probe," in *2018 3rd Biennial South African Biomedical Engineering Conference (SAIBMEC)*. Stellenbosch: IEEE, Apr. 2018, pp. 1–5.

[92] T. Röddiger, D. Wolffram, D. Laubenstein, M. Budde, and M. Beigl, "Towards Respiration Rate Monitoring Using an In-Ear Headphone Inertial Measurement Unit," in *Proceedings of the 1st International Workshop on Earable Computing*. London United Kingdom: ACM, Sep. 2019, pp. 48–53.

[93] M. M. Rahman, T. Ahmed, M. Y. Ahmed, E. Nemati, M. Dinh, N. Folkman, M. M. Hasan, J. Kuang, and J. A. Gao, "Towards Motion-Aware Passive Resting Respiratory Rate Monitoring Using Earbuds," in *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. Athens, Greece: IEEE, Jul. 2021, pp. 1–4.

[94] T. Ahmed, M. M. Rahman, M. Yusuf Ahmed, E. Nemati, M. Dinh, N. Folkman, J. Kuang, and A. Gao, "RRMonitor: A Resource-Aware End-to-End System for Continuous Monitoring of Respiration Rate Using Earbuds," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Mexico: IEEE, Nov. 2021, pp. 2463–2467.

[95] A. Kumar, V. Mitra, C. Oliver, A. Ullal, M. Biddulph, and I. Mance, "Estimating Respiratory Rate From Breath Audio Obtained Through Wearable Microphones," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Mexico: IEEE, Nov. 2021, pp. 7310–7315.

[96] T. Ahmed, M. M. Rahman, E. Nemati, M. Y. Ahmed, J. Kuang, and A. J. Gao, "Remote Breathing Rate Tracking in Stationary Position Using the Motion and Acoustic Sensors of Earables," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, pp. 1–22.

[97] T. Hao, G. Xing, and G. Zhou, "RunBuddy: A smartphone system for running rhythm monitoring," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*. Osaka, Japan: ACM Press, 2015, pp. 133–144.

[98] F. Gu, J. Niu, S. K. Das, Z. He, and X. Jin, "Detecting breathing frequency and maintaining a proper running rhythm," *Pervasive and Mobile Computing*, vol. 42, pp. 498–512, 2017.

[99] G. Pressler, J. Mansfield, H. Pasterkamp, and G. Wodicka, "Detection of Respiratory Sounds at the External Ear," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 12, pp. 2089–2096, Dec. 2004.

[100] E. P. Widmaier, H. Raff, A. J. Vander, and K. T. Strang, *Vander's Human Physiology: The Mechanisms of Body Function*. McGraw-Hill, 2011.

[101] J. D. Pollock and A. N. Makaryus, *Physiology, Cardiac Cycle*. StatPearls Publishing, Treasure Island (FL), 2023.

[102] W. Commons, "A wiggers diagram, showing the cardiac cycle events occuring in the left ventricle." 2016, [Accessed 21-08-24]. [Online]. Available: https://commons.wikimedia.org/wiki/File:Wiggers_Diagram_2.svg

[103] Z. S. Bruss and A. Raja, *Physiology, Stroke Volume*. StatPearls Publishing, Treasure Island (FL), 2023.

[104] D. Yazdi, S. Sridaran, S. Smith, C. Centen, S. Patel, E. Wilson, L. Gillon, S. Kapur, J. A. Tracy, K. Lewine, D. M. Systrom, and C. A. MacRae, "Noninvasive Scale Measurement of Stroke Volume and Cardiac Output Compared With the Direct Fick Method: A Feasibility Study," *Journal of the American Heart Association*, vol. 10, no. 24, p. e021893, Dec. 2021.

[105] A. N. De Maria and A. Raisinghani, "Comparative Overview of Cardiac Output Measurement Methods: Has Impedance Cardiography Come of Age?" *Congestive Heart Failure*, vol. 6, no. 2, pp. 60–73, 2000.

[106] W. H. Fares, S. K. Blanchard, G. A. Stouffer, P. P. Chang, W. D. Rosamond, H. J. Ford, and R. M. Aris, "Thermodilution and Fick cardiac outputs differ: Impact on pulmonary hypertension evaluation," *Canadian Respiratory Journal : Journal of the Canadian Thoracic Society*, vol. 19, no. 4, p. 261, 2012-07/2012-08.

[107] "Finapres; NOVA."

[108] B. Bein and J. Renner, "Best practice & research clinical anaesthesiology: Advances in haemodynamic monitoring for the perioperative patient: Perioperative cardiac output monitoring," *Best Practice & Research Clinical Anaesthesiology*, vol. 33, no. 2, pp. 139–153, Jun. 2019.

[109] R. Couceiro, P. Carvalho, R. P. Paiva, J. Henriques, M. Antunes, I. Quintal, and J. Muehlsteff, "Beat-to-beat cardiac output inference using heart sounds," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Boston, MA: IEEE, Aug. 2011, pp. 5657–5661.

[110] R. P. Paiva, P. Carvalho, R. Couceiro, J. Henriques, M. Antunes, I. Quintal, and J. Muehlsteff, "Beat-to-beat systolic time-interval measurement from heart sounds and ECG," *Physiological Measurement*, vol. 33, no. 2, pp. 177–194, Feb. 2012.

[111] Y. D. Shin, K. H. Yim, S. H. Park, Y. W. Jeon, J. H. B. Bae, T. S. Lee, M. H. Kim, and Y. J. Choi, "The First Heart Sound Amplitude Using Digital Esophageal Stethoscope System is Proportional to The Change in Cardiac Output," *Pakistan Journal of Medical Sciences*, vol. 30, no. 2, Dec. 1969.

[112] K. H. Wesseling, J. R. Jansen, J. J. Settels, and J. J. Schreuder, "Computation of aortic flow from pressure in humans using a nonlinear, three-element model," *Journal of Applied Physiology*, May 1993.

[113] Q. Y. Lee, S. J. Redmond, G. S. Chan, P. M. Middleton, E. Steel, P. Malouf, C. Critoph, G. Flynn, E. O'Lone, and N. H. Lovell, "Estimation of cardiac output and systemic vascular resistance using a multivariate regression model with features selected from the finger photoplethysmogram and routine cardiovascular measurements," *BioMedical Engineering OnLine*, vol. 12, no. 1, p. 19, 2013.

[114] L. M. Johnson, S. L. Swarner, A. van der Straten, and G. D. Rothrock, *Methods for Assessing the Adherence to Medical Devices*, ser. RTI Press Methods Report Series. Research Triangle Park (NC): RTI Press, 2016.

[115] D. Nachman, K. Constantini, G. Poris, L. Wagnert-Avraham, S. D. Gertz, R. Littman, E. Kabakov, A. Eisenkraft, and Y. Gepner, "Wireless, non-invasive, wearable device for continuous remote monitoring of hemodynamic parameters in a swine model of controlled hemorrhagic shock," *Scientific Reports*, vol. 10, no. 1, p. 17684, Oct. 2020.

[116] M. M. Milani, P. E. Abas, and L. C. De Silva, "A critical review of heart sound signal segmentation algorithms," *Smart Health*, vol. 24, p. 100283, 2022.

[117] P. Sharma, S. A. Imtiaz, and E. Rodriguez-Villegas, "An Algorithm for Heart Rate Extraction From Acoustic Recordings at the Neck," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 246–256, Jan. 2019.

[118] P. White, "TRANSFORMS, WAVELETS," in *Encyclopedia of Vibration*, S. Braun, Ed. Oxford: Elsevier, Jan. 2001, pp. 1419–1435.

[119] M. N. Ali, E.-S. A. El-Dahshan, and A. H. Yahia, "Denoising of Heart Sound Signals Using Discrete Wavelet Transform," *Circuits, Systems, and Signal Processing*, vol. 36, no. 11, pp. 4482–4497, Nov. 2017.

[120] J. Barrios-Muriel, F. Romero, F. J. Alonso, and K. Gianikellis, "A simple SSA-based de-Noising technique to remove ECG interference in EMG signals," *Biomedical Signal Processing and Control*, vol. 30, pp. 117–126, Sep. 2016.

[121] F. Ghaderi, H. R. Mohseni, and S. Sanei, "Localizing heart sounds in respiratory signals using singular spectrum analysis," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12, pp. 3360–3367, 2011.

[122] Y. Diao, Y. Ma, D. Xu, W. Chen, and Y. Wang, "A novel gait parameter estimation method for healthy adults and postoperative patients with an ear-worn sensor," *Physiological Measurement*, vol. 41, no. 5, p. 05NT01, Jun. 2020.

[123] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[124] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2021.

[125] C. Yu, R. E. Zezario, S.-S. Wang, J. Sherman, Y.-Y. Hsieh, X. Lu, H.-M. Wang, and Y. Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *arXiv preprint arXiv:2001.01538*, 2020.

[126] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, vol. 9351, pp. 234–241.

[127] E. J. Nustede and J. Anemüller, "Towards speech enhancement using a variational U-Net architecture," Mar. 2021.

[128] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer unet for image denoising," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, May 2022.

[129] V. S. Murahari and T. Ploetz, "On Attention Models for Human Activity Recognition," *arXiv:1805.07648 [cs]*, May 2018.

[130] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," Dec. 2021.

[131] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, and Y. Gu, "A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations," *Expert Systems with Applications*, vol. 242, p. 122807, May 2024.

[132] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked Autoencoders that Listen," Jan. 2023.

[133] J. Ahn, H.-K. Ra, H. J. Yoon, S. H. Son, and J. Ko, "On-device filter design for self-identifying inaccurate heart rate readings on wrist-worn PPG sensors," *IEEE access : practical innovations, open solutions*, vol. 8, pp. 184 774–184 784, 2020.

[134] V. Goverdovsky, W. Von Rosenberg, T. Nakamura, D. Looney, D. J. Sharp, C. Papavassiliou, M. J. Morrell, and D. P. Mandic, "Hearables: Multimodal physiological in-ear sensing," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[135] MP. Murray, GB. Spurr, SB. Sepic, GM. Gardner, and LA. Mollinger, "Treadmill vs. floor walking: Kinematics, electromyogram, and heart rate," *Journal of applied physiology*, vol. 59, no. 1, pp. 87–91, 1985.

[136] S. Passler, N. Müller, and V. Senner, "In-ear pulse rate measurement: A valid alternative to heart rate derived from electrocardiography?" *Sensors*, vol. 19, no. 17, p. 3641, 2019.

[137] J. A. Patterson, D. C. McIlwraith, and G.-Z. 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*. IEEE, 2009, pp. 286–291.

[138] Consumer Technology Association, "Physical Activity Monitoring for Heart Rate ANSI/CTA-2065," 2018.

[139] *Zero-Height SiSonic Microphone with Extended Low Frequency Performance*, Mar. 2013, rev. D.

[140] *Single-Supply Integrated Optical Module for HR and SpO2 Measurement*, Mar. 2019, rev. 0.

[141] MA. Ali and PM. Shemi, "An improved method of audio denoising based on wavelet transform," in *2015 International Conference on Power, Instrumentation, Control and Computing*. IEEE, 2015, pp. 1–6.

[142] M. N. Ali, E.-S. A. El-Dahshan, and A. H. Yahia, "Denoising of heart sound signals using discrete wavelet transform," *Circuits, Systems, and Signal Processing*, vol. 36, no. 11, pp. 4482–4497, 2017.

[143] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[144] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 241–246.

[145] M. T. Nguyen, W. W. Lin, and J. H. Huang, "Heart Sound Classification Using Deep Learning Techniques Based on Log-mel Spectrogram," *Circuits, Systems, and Signal Processing*, Aug. 2022.

[146] F. Demir, A. Şengür, V. Bajaj, and K. Polat, "Towards the classification of heart sounds based on convolutional deep neural network," *Health Information Science and Systems*, vol. 7, no. 1, p. 16, Aug. 2019.

[147] W. Xu, X. Deng, S. Guo, J. Chen, L. Sun, X. Zheng, Y. Xiong, Y. Shen, and X. Wang, "High-Resolution U-Net: Preserving Image Details for Cultivated Land Extraction," *Sensors (Basel, Switzerland)*, vol. 20, no. 15, p. 4064, Jul. 2020.

[148] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (R2U-net) for medical image segmentation," 2018.

[149] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[150] Q. Man, Y.-I. Cho, S.-G. Jang, and H.-J. Lee, "Transformer-based GAN for new hairstyle generative networks," *Electronics*, vol. 11, no. 2106, 2022.

[151] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, "The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results."

[152] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss Functions for Neural Networks for Image Processing," Apr. 2018.

[153] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4.

[154] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiological Measurement*, vol. 29, no. 1, pp. 15–32, Jan. 2008.

[155] "SPU1410LR5H-QB," https://www.digikey.co.uk/en/products/\detail/knowles/SPU1410LR5H-QB/3621629, 2023.

[156] HiFiBerry, "HiFiBerry DAC+ ADC Pro | HiFiBerry."

[157] D. Giavarina, "Understanding bland altman analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[158] C. Massaroni, A. Nicolò, D. Lo Presti, M. Sacchetti, S. Silvestri, and E. Schena, "Contact-Based Methods for Measuring Respiratory Rate," *Sensors*, vol. 19, no. 4, p. 908, Feb. 2019.

[159] C. Hu, T. Kandappu, Y. Liu, C. Mascolo, and D. Ma, "BreathPro: Monitoring breathing mode during running with earables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–25, 2024.

[160] J. B. Elsner and A. A. Tsonis, *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Springer Science & Business Media, 1996.

[161] E. R. Castillo and D. E. Lieberman, "Shock attenuation in the human lumbar spine during walking and running," *Journal of Experimental Biology*, vol. 221, no. 9, p. jeb177949, 2018.

[162] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, and C. Mascolo, "hEARt: Motion-resilient heart rate monitoring with in-ear microphones," in *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2023, pp. 200–209.

[163] D. Ma, A. Ferlini, and C. Mascolo, "OESense: Employing occlusion effect for in-Ear human sensing," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. Virtual Event Wisconsin: ACM, Jun. 2021, pp. 175–187.

[164] F. Yasuma and J.-i. Hayano, "Respiratory sinus arrhythmia: Why does the heartbeat synchronize with respiratory rhythm?" *Chest*, vol. 125, no. 2, pp. 683–690, 2004.

[165] B. Aysin and E. Aysin, "Effect of respiration in heart rate variability (HRV) analysis," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 1776–1779.

[166] G. Blain, O. Meste, and S. Bermon, "Influences of breathing patterns on respiratory sinus arrhythmia in humans during exercise," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 288, no. 2, pp. H887–H895, Feb. 2005.

[167] C. P. Hoffmann, G. Torregrosa, and B. G. Bardy, "Sound stabilizes locomotor-respiratory coupling and reduces energy cost," *PloS One*, vol. 7, no. 9, p. e45206, 2012.

[168] J.-F. Gariépy, K. Missaghi, and R. Dubuc, "The interactions between locomotion and respiration," in *Progress in Brain Research*. Elsevier, 2010, vol. 187, pp. 173–188.

[169] C. R. Illidi, J. Stang, J. Melau, J. Hisdal, and T. Stensrud, "Does Cold-Water Endurance Swimming Affect Pulmonary Function in Healthy Adults?" *Sports*, vol. 9, no. 1, p. 7, Jan. 2021.

[170] D. A. Mahler, B. Hunter, T. Lentine, and J. Ward, "Locomotor-respiratory coupling develops in novice female rowers with training," *Medicine and Science in Sports and Exercise*, vol. 23, no. 12, pp. 1362–1366, Dec. 1991.

[171] J. O'Halloran, J. Hamill, W. J. McDermott, J. G. Remelius, and R. E. A. Van Emmerik, "Locomotor-respiratory coupling patterns and oxygen consumption during walking above and below preferred stride frequency," *European Journal of Applied Physiology*, vol. 112, no. 3, pp. 929–940, Mar. 2012.

[172] M. A. Daley, D. M. Bramble, and D. R. Carrier, "Impact Loading and Locomotor-Respiratory Coordination Significantly Influence Breathing Dynamics in Running Humans," *PLoS ONE*, vol. 8, no. 8, p. e70752, Aug. 2013.

[173] C. P. Hoffmann, G. Torregrosa, and B. G. Bardy, "Sound Stabilizes Locomotor-Respiratory Coupling and Reduces Energy Cost," *PLoS ONE*, vol. 7, no. 9, p. e45206, Sep. 2012.

[174] J. Morales, J. Moeyersons, P. Armanac, M. Orini, L. Faes, S. Overeem, M. Van Gilst, J. Van Dijk, S. Van Huffel, R. Bailon *et al.*, "Model-based evaluation of methods for respiratory sinus arrhythmia estimation," *IEEE transactions on biomedical engineering*, vol. 68, no. 6, pp. 1882–1893, 2020.

[175] S. Fuchs and A. Rochet-Capellan, "The respiratory foundations of spoken language," *Annual Review of Linguistics*, vol. 7, pp. 13–30, 2021.

[176] J. Cioffi and T. Kailath, "Fast, recursive-least-squares transversal filters for adaptive filtering," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 304–337, 1984.

[177] "Normal respiration rate: For adults and all ages, and how to measure," Feb. 2019.

[178] A. Oliveira and A. Marques, "Respiratory sounds in healthy people: A systematic review," *Respiratory medicine*, vol. 108, no. 4, pp. 550–570, 2014.

[179] J. O'Halloran, J. Hamill, W. J. McDermott, J. G. Remelius, and R. E. Van Emmerik, "Locomotor-respiratory coupling patterns and oxygen consumption during walking above and below preferred stride frequency," *European Journal of Applied Physiology*, vol. 112, no. 3, pp. 929–940, 2012.

[180] M. A. Daley, D. M. Bramble, and D. R. Carrier, "Impact loading and locomotor-respiratory coordination significantly influence breathing dynamics in running humans," *PloS one*, vol. 8, no. 8, p. e70752, 2013.

[181] A. Arefeen, A. Akbari, S. I. Mirzadeh, R. Jafari, B. A. Shirazi, and H. Ghasemzadeh, "Inter-beat interval estimation with tiramisu model: A novel approach with reduced error," 2021.

[182] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[183] M. Prabhu, "Cardiac output measurement," *Anaesthesia & Intensive Care Medicine*, vol. 8, no. 2, pp. 63–66, Feb. 2007.

[184] C. A. Vella and R. A. Robergs, "A review of the stroke volume response to upright exercise in healthy subjects," *British Journal of Sports Medicine*, vol. 39, no. 4, pp. 190–195, Apr. 2005.

[185] P. Loprinzi, S. Maskalick, and V. Veigl, "Chapter 5 - exercise physiology," in *Orthopaedic Physical Therapy Secrets (Third Edition)*, 3rd ed., J. D. Placzek and D. A. Boyce, Eds. Elsevier, 2017, pp. 35–43.

[186] M. De Marco, E. Gerdts, C. Mancusi, M. J. Roman, M. T. Lønnebakken, E. T. Lee, B. V. Howard, R. B. Devereux, and G. de Simone, "Influence of Left Ventricular Stroke Volume on Incident Heart Failure in a Population with Preserved Ejection Fraction (From the Strong Heart Study)," *The American journal of cardiology*, vol. 119, no. 7, pp. 1047–1052, Apr. 2017.

[187] L. A. H. Critchley, "Evaluation of a cardiac output monitor," in *Perioperative Hemodynamic Monitoring and Goal Directed Therapy: From Theory to Practice*, M. Cannesson and R. Pearse, Eds. Cambridge: Cambridge University Press, 2014, pp. 120–131.

[188] B. Gårdbæk and P. Kidmose, "On the Origin of Cardiovascular Sounds Recorded from the Ear," Jun. 2024.

[189] X. Xu, Q. Tang, and Z. Chen, "Improved U-Net Model to Estimate Cardiac Output Based on Photoplethysmography and Arterial Pressure Waveform," *Sensors*, vol. 23, no. 22, p. 9057, Nov. 2023.

[190] B. Schuller, "Audio features," in *Intelligent Audio Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 41–97.

[191] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, "Deep learning in biomedicine," *Nature Biotechnology*, vol. 36, no. 9, pp. 829–838, Oct. 2018.

[192] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[193] S. A. A. Ahmed, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, "ASiT: Local-global audio spectrogram vision transformer for event classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3684–3693, 2024.

[194] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Y. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan, "Self-Supervised Learning for Time Series Analysis: Taxonomy, Progress, and Prospects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6775–6794, Oct. 2024.

[195] Z. Liu, A. Alavi, M. Li, and X. Zhang, "Self-supervised learning for time series: Contrastive or generative?" 2024.

[196] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[197] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[198] J. Waters, J. Stuchbury-Wass, Y. Liu, K.-J. Butkow, and C. Mascolo, "Deep-learning based segmentation of in-ear cardiac sounds," 2024.

[199] Y. Liu, Q. Yang, K.-J. Butkow, J. Stuchbury-Wass, D. Ma, and C. Mascolo, "Tidal volume estimation dataset," 2024.

[200] L. A. Critchley and J. A. Critchley, "A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques," *Journal of Clinical Monitoring and Computing*, vol. 15, no. 2, pp. 85–91, Feb. 1999.

[201] H. N. Patel, T. Miyoshi, K. Addetia, M. P. Henry, R. Citro, M. Daimon, P. Gutierrez Fajardo, R. R. Kasliwal, J. N. Kirkpatrick, M. J. Monaghan, D. Muraru, K. O. Ogunyankin, S. W. Park, R. E. Ronderos, A. Sadeghpour, G. M. Scalia, M. Takeuchi, W. Tsang, E. S. Tucay, A. C. Tude Rodrigues, A. Vivekanandan, Y. Zhang, M. Schreckenberg, M. Blankenhagen, M. Degel, A. Rossmanith, V. Mor-Avi, F. M. Asch, R. M. Lang, A. D. Prado, E. Filipini, A. Kwon, S. Hoschke-Edwards, T. R. Afonso, B. Thampinathan, M. Sooriyakanthan, T. Zhu, Z. Wang, Y. Wang, L. Yin, S. Li, R. Alagesan, S. Balasubramanian, R. V. A. Ananth, M. Bansal, A. Alizadehasl, L. Badano, E. Bossone, D. Di Vece, M. Bellino, T. Nakao, T. Kawata, M. Hirokawa, N. Sawada, Y. Nabeshima, H. R. Yun, and J.-w. Hwang, "Normal Values of Cardiac Output and Stroke Volume According to Measurement Technique, Age, Sex, and Ethnicity: Results of the World Alliance of Societies of Echocardiography Study," *Journal of the American Society of Echocardiography*, vol. 34, no. 10, pp. 1077–1085.e1, Oct. 2021.

[202] J. M. Evans, S. Wang, C. Greb, V. Kostas, C. F. Knapp, Q. Zhang, E. S. Roemmele, M. B. Stenger, and D. C. Randall, "Body Size Predicts Cardiac and Vascular Resistance Effects on Men's and Women's Blood Pressure," *Frontiers in Physiology*, vol. 8, p. 561, Aug. 2017.

[203] T. Collis, R. B. Devereux, M. J. Roman, G. de Simone, J.-L. Yeh, B. V. Howard, R. R. Fabsitz, and T. K. Welty, "Relations of Stroke Volume and Cardiac Output to Body Composition," *Circulation*, vol. 103, no. 6, pp. 820–825, Feb. 2001.

[204] B. U. Demirel, T. Dang, K. Al-Naimi, F. Kawsar, and A. Montanari, "Unobtrusive Air Leakage Estimation for Earables with In-ear Microphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, pp. 1–29, Dec. 2023.

[205] T. Röddiger, T. King, D. R. Roodt, C. Clarke, and M. Beigl, "OpenEarable: Open hardware earable sensing platform," in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. New York, NY, USA: Association for Computing Machinery, 2023.

Appendix A

Hardware Design

To overcome the lack of API access to internally facing microphones on commercial earables, we designed a custom earable prototype which was used for all the data collection exercises described in this thesis. In this section, we detail the design of the earable hardware.

A.1 Ear-hook

Using CAD software, we designed an earhook which housed the internally facing microphone and the speaker. The shape of the earhook, which can be seen in Figure 3.4.1, was based on the design of Röddiger et al. [205]. In component (d) of Figure 3.4.1 (the 3D printed front cap), we adhered a Knowles SPU1410LR5H-QB microphone with the diaphragm facing towards the ear tip using superglue. We then glued a speaker (a standard true wireless stereo 8mm, 16 Ohm earbud speaker) behind the microphone in the same manner. The wiring from the two components was soldered into a female RJ45 connector for ease of connecting the earbuds to the Raspberry Pi board. We plugged a lightweight ethernet cable into the connector. The left and right ear-hooks are mirror images of one another. The circuit diagrams of the microphone and speaker connections are provided in Figure A.2.1.

A.2 Custom PCB

We designed a custom PCB to interface between the earbuds and the Raspberry Pi audio codec. Figure A.2.1 provides the circuit layout of this board, and the PCB layout is shown in Figure A.2.2. The PCB contains two female RJ45 connectors that interface with the

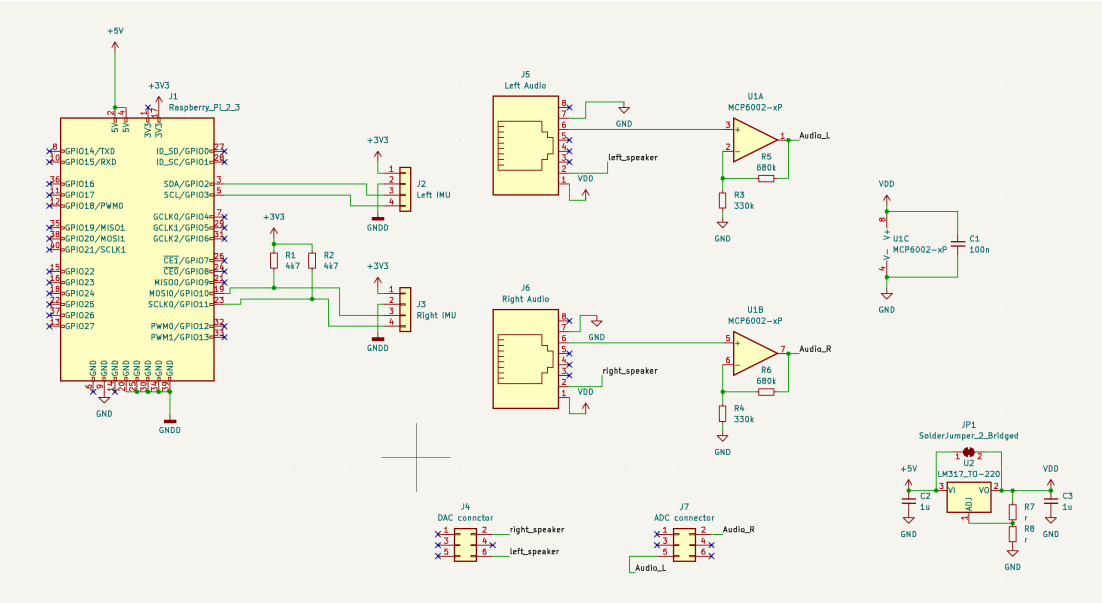


Figure A.2.1: Circuit diagram of the custom earbud.

earbuds' ethernet cables. The PCB also contains one MCP6004 non-inverting operational amplifier per channel (earbud). These amplifiers are controlled by potentiometers, allowing for adjustable gain if required. The gain value was set to unity for all data collection exercises. The PCB was designed to interface with the HiFiBerry DAC+ ADC pro audio codec and so relays the amplified microphone signals to the analog inputs of the codec.

A.3 Data sampling

The data was sampled using a Python script on a Raspberry Pi 4. With this script, the sampling rate could easily be adjusted to fit the application and data was written directly onto the Raspberry Pi 4. Data was transferred from the device onto the local computer using *SCP*, and processing was done offline unless specified otherwise.

During the data collection exercise, the investigator accessed the Raspberry Pi using SSH and ran the script to record from the microphones remotely.

A.4 Wearing the device

As shown in Figure 3.4.1(b), the earable was powered by a portable power bank and all circuitry was placed into a chest strap to maintain freedom of movement of the participant.

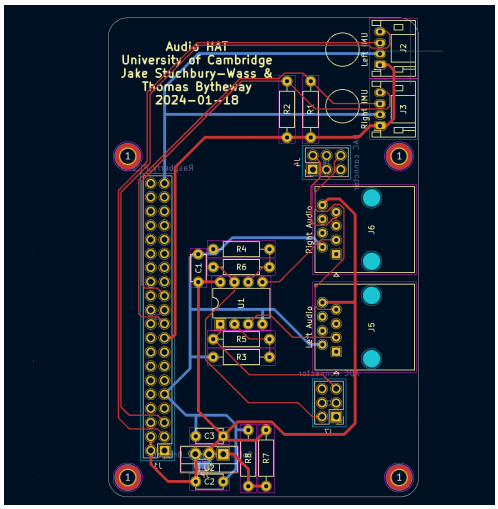


Figure A.2.2: PCB layout of the custom board used to interface the earbuds with the Raspberry Pi

During walking and running, we secured the chest bag to the participant using a velcro strap to prevent it from bouncing while undergoing motion, thus limiting motion artefacts in the signal caused by the movement of the wires.

Participants were asked to insert the earbuds into their ears and adjust the earbuds until a tight seal was created. The quality of the signal was visually assessed by the investigator before beginning the data collection exercise. To prevent the earbuds from falling out of the participants' ears, a sports headband was placed over the earbuds. The use of the chest bag and headband were necessary due to the research-grade earbud prototype, but would not be required if the algorithms were integrated into commercial devices.