5-2024

An evaluation of heart rate monitoring with in-ear microphones under motion

Kayla-Jade BUTKOW

Ting DANG

Andrea FERLINI

Dong MA
Singapore Management University, dongma@smu.edu.sg

Yang LIU

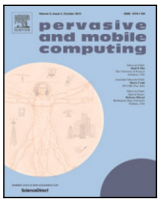
See next page for additional authors

Pervasive and Mobile Computing 100 (2024) 101913

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc



An evaluation of heart rate monitoring with in-ear microphones under motion

Kayla-Jade Butkow [a,*], Ting Dang [a], Andrea Ferlini [a], Dong Ma [b], Yang Liu [a], Cecilia Mascolo [a]

[a] University of Cambridge, Cambridge, UK
[b] Singapore Management University, Singapore

ARTICLE INFO

Keywords:
Earable
Heart rate
Motion artefact
In-ear audio

ABSTRACT

With the soaring adoption of in-ear wearables, the research community has started investigating suitable in-ear heart rate detection systems. Heart rate is a key physiological marker of cardiovascular health and physical fitness. Continuous and reliable heart rate monitoring with wearable devices has therefore gained increasing attention in recent years. Existing heart rate detection systems in wearables mainly rely on photoplethysmography (PPG) sensors, however, these are notorious for poor performance in the presence of human motion. In this work, leveraging the occlusion effect that enhances low-frequency bone-conducted sounds in the ear canal, we investigate for the first time *in-ear audio-based motion-resilient* heart rate monitoring. We first collected heart rate-induced sounds in the ear canal using an in-ear microphone under seven stationary activities and two full-body motion activities (i.e., walking, and running). Then, we devised a novel deep learning based motion artefact (MA) mitigation framework to denoise the in-ear audio signals, followed by a heart rate estimation algorithm to extract heart rate. With data collected from 15 subjects over nine activities, we demonstrate that hEARt, our end-to-end approach, achieves a mean absolute error (MAE) of 1.88 ± 2.89 BPM, 6.83 ± 5.05 BPM, and 13.19 ± 11.37 BPM for stationary, walking, and running, respectively, opening the door to a new non-invasive and affordable heart rate monitoring with useable performance for daily activities. Not only does hEARt outperform previous in-ear heart rate monitoring work, but it outperforms reported in-ear PPG performance.

1. Introduction

Heart rate (HR) is an excellent indicator of fitness level, and is strongly associated with cardiovascular disease and mortality risk. HR monitoring can help design workout routines to maximize 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, the change in time between successive beats, is a predictor of physical and mental health. Heart rate variability, a proxy for autonomic nervous system behaviour, is predictive of aerobic fitness when measured during both maximal and sub-maximal exercise [1]. Thus measuring HR under motion is critical for monitoring human health and wellbeing.

Electrocardiographic (ECG) telemetry monitoring is the standard for HR and heart rate variability monitoring. However ECGs need to be connected to the body with leads making them unsuitable for realistic and mobile settings. Although attempts to devise portable ECG, such as ECG chest straps, have been introduced, they remain cumbersome, uncomfortable, and inconvenient. New

* Corresponding author.
E-mail address: kjb85@cam.ac.uk (K.-J. Butkow).

https://doi.org/10.1016/j.pmcj.2024.101913
Received 10 August 2023; Received in revised form 12 February 2024; Accepted 6 March 2024
Available online 7 March 2024

smartwatches include a single-lead ECG, however to use them, the user is required to remain still and to close the ECG circuit with their fingers. They are thus unable to monitor continuously.

Recent trends in wearables have led to a proliferation of studies investigating different sensors on smartwatches, earables, and other wearables for HR monitoring. Photoplethysmography (PPG) sensors, which measure light scatter as a result of blood flow, are most commonly adopted due to their non-invasiveness, easy implementation and low cost. Although PPG is effective and accurate for HR measurements under stationary conditions [2], it is sensitive to motion artefacts (MAs) caused by users' body movement or physical activities [2–4]. Due to these MAs, 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 [2–4], a consensus on the best commercially available device to monitor the wearer's HR whenever motion is concerned, is yet to be found. Moreover, intense motion, like walking and running, yields substantial deviations from ground-truth (GT), resulting in average errors up to 30% across a wide-spectrum of wrist-worn devices [2]. Dealing with interference from MAs is thus an open and challenging problem in HR estimation.

Due to the limitations of wrist-based PPG, researchers have started investigating alternative wearables for HR monitoring under motion. With the rapid spreading of ear-worn wearables (earables) in daily life [5], earables can be a portable and non-invasive means of continuous HR detection. Particularly, due to their pervasiveness during physical activity (specifically while walking and running), the earable form factor can be exploited for HR monitoring while under motion. Research has started to emerge in earable-based PPG for continuous HR sensing [6]. However, despite being a promising modality, real world performance of earable PPG under motion is still poor [3,7]. Indeed, similar to what is observed for wrist-worn devices [2], errors around 30% have been reported [7].

Current commercial earables are equipped with multiple sensors, including outer and inner ear microphones which fulfil fundamental functionalities of the device (e.g., speech detection and active noise cancellation). Recently, Martin and Voix [8] proposed to measure HR using a microphone placed in the human ear canal. When the ear canal opening is sealed by the earbuds, the cavity formed between the ear tip and eardrum enables an enhancement of low-frequency sounds, called the occlusion effect [9]. As a result, heartbeat-induced sounds that propagate to the ear canal through bone conduction are amplified and can be leveraged for HR estimation. Their results show an error of 5.6% for HR determination under stationary conditions. However, [8] only demonstrated the feasibility of measuring HR with in-ear microphones while an individual is stationary: *how in-ear microphone HR measurement performs under active scenarios remains unclear and unexplored*.

In this work, we focus on in-ear HR estimation under both stationary and active scenarios (i.e. walking, and running). The biggest hurdle to accurate HR measurement is motion-induced interference, which is amplified by the occlusion effect along with the heart sounds [10]. 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 HR and walking frequency (i.e. cadence) are similar (both around 1.5–2.3 Hz [11]), it is hard to separate them in the frequency domain.

To address these challenges, we propose a processing pipeline for accurate HR detection in the presence of MAs, namely, walking and running. Different from previous audio-based HR estimation works [6,8,12,13], we also validate the functioning of our technique in the presence of music, showing that the proposed approach can determine HR 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 HR estimation under motion cases with good performance. Specifically, with mean absolute percentage error (MAPE) less than 10% while stationary, walking and running, this system is accurate according to ANSI standards for HR accuracy for a physical monitoring device [2,14]. Additionally, because of the artifacts considered, the vantage points (the ears), and the device form-factor (earables), our work is directly comparable to [7]. Notably, we significantly outperform in-ear PPG [7] (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 [15,16] and affordable offering additional appeal for continuous HR estimation.

The contribution of this work can be summarized as follows: **(i)** We explore HR estimation with in-ear microphones and present an analysis of the interference imposed by full body motion. **(ii)** We propose a novel pipeline for HR estimation under MAs, consisting of a CNN-based module using U-Net architecture to enhance audio-based heart sounds with ECG as a reference, and an estimation module using signal processing to estimate HR from cleaned signals. We further leverage transfer learning that pre-trains the model using a large heart sounds dataset and fine-tunes 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 audio-based heart sounds captured by earables using ECG signals. **(iii)** We built a custom earbud prototype and collected data from 15 subjects. Results show that we can achieve mean absolute error (MAE) of 1.88 ± 2.89 BPM, 6.83 ± 5.05 BPM, and 13.19 ± 11.37 BPM for stationary, walking, and running, respectively, demonstrating the effectiveness of the proposed approach in combating MAs.

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 HR estimation under motion conditions.

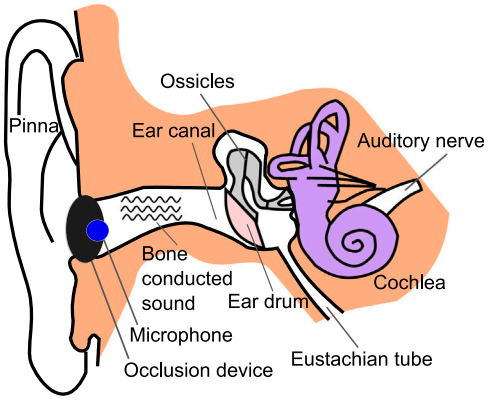


Fig. 1. The occlusion effect and the anatomy of the ear.

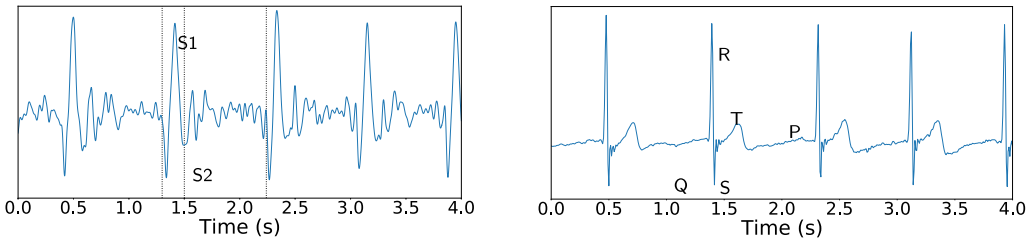


Fig. 2. The (left) sound signal captured by the internal microphone indicating the S1 and S2 heart sounds, and the (right) corresponding ECG signal showing the three main components of the ECG (QRS complex (ventricular depolarization), *T* wave (ventricular repolarization), P wave (atrial depolarization)).

2.1. In-ear heart sound acquisition

Bone conduction, a physiological phenomenon whereby sound is conducted through the bones directly to the inner ear, causes vibrations in the walls of the ear [9]. When the ear canal is occluded, the increase in impedance at the entrance of the ear canal results in an amplification of low frequency sounds conducted by the bones [9]. This effect, illustrated in Fig. 1, is known as the occlusion effect. Since bone conveys low-frequency sounds [17], the bone-conducted heart sounds are amplified in the occluded ear canal [8]. Heart sounds can thus be detected using a microphone placed inside the occluded ear canal. An example showing the heart sounds captured by the internal microphone is shown in Fig. 2. Clearly, 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 HR monitoring. The correlation between the in-ear captured audio and the ECG signal is also evident in Fig. 2.

2.2. Motion artefacts analysis and challenges

In-ear microphone based HR estimation suffers from human MAs since the occlusion effect not only amplifies the heartbeat-induced sound, but also enhances other bone-conducted sounds and vibrations inside the body [10,18]. Fig. 3(a–c) illustrates the recorded audio signals from the in-ear microphone while stationary, walking and running within a seven second window. Fig. 3(d–e) are spectrograms of the activities shown over a longer timescale so that trends in HR can be seen. The heartbeat is clearly observable when an individual is stationary in Fig. 3(a), with HR frequency lying around 1 Hz and its 1st and 2nd harmonics clearly observable in Fig. 3(d). Contrastingly, it is completely overwhelmed by the amplified step sounds in Fig. 3(b) (note the different scales of the y-axis), with the periodic peaks corresponding to the sound of foot strikes that propagate through the human skeleton, resulting in a significantly higher energy observed around 1.7 Hz (the cadence) in Fig. 3(e). Here, we can no longer see the HR or its harmonics, as they are overwhelmed by the walking. In this case, the periods of the heart sounds and walking are similar, resulting in an overlap in the frequency domain, making it challenging to split the heart sounds and walking signals and estimate the HR either in time domain or frequency domain.

The heartbeats are further affected and obscured by foot strikes while running (Fig. 3(c)) which exhibit far stronger energy than any of the other activities, with high energy at 2.8 Hz (Fig. 3(f)) again corresponding to the cadence. Here the low-amplitude heart sounds are entirely buried in the high-amplitude step sounds.

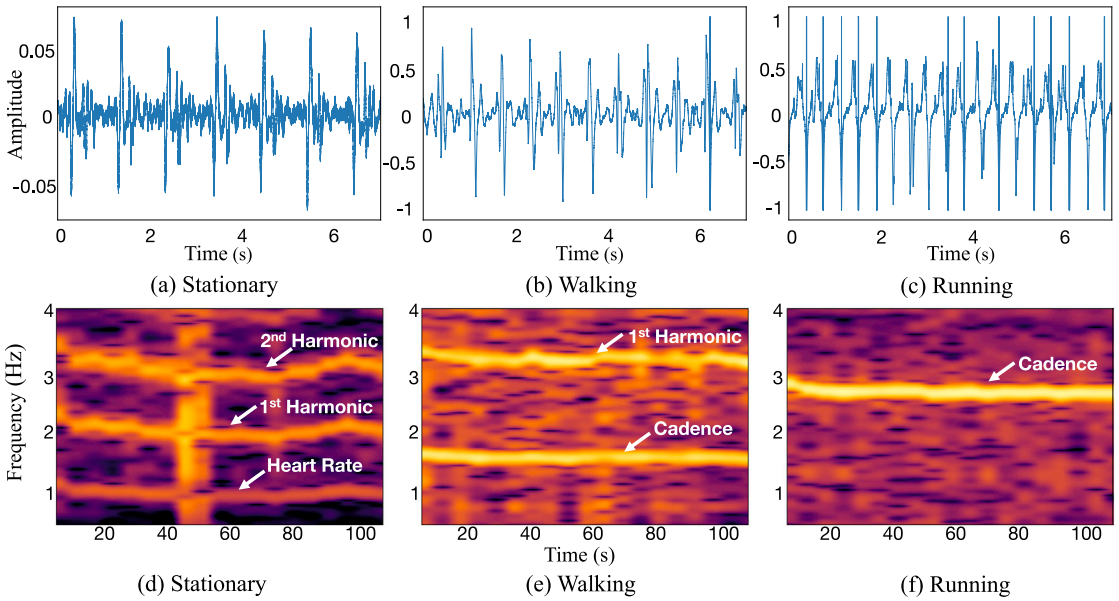


Fig. 3. Time domain representations and spectrograms of audio signals captured by the in-ear microphone.

3. Motion-resilient HR estimation

Typical signal processing techniques have shown effectiveness in HR estimation in the stationary case [8]. However, they do not adequately isolate the heart sounds from the corrupted audio under MAs. As previously discussed, 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 [19,20]. 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 (DL) for denoising in numerous fields [21,22], we propose a novel pipeline using DL to eliminate MAs in audio and estimate HR. In the following sections, we first present a signal processing approach for HR estimation, and then the proposed DL pipeline for MA removal.

3.1. Signal processing for HR estimation

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

First, we compute the Hilbert transform of the audio to calculate the HR envelope. We then compute the spectrum of the envelope using Fast Fourier Transform (FFT) and detect the dominant peaks which are converted to the HR. This approach shows good performance on a clean and stationary signal (see Section 5.2). However, when audio signals are corrupted with motions, dominant peaks in the spectrum correspond to motions, rather than the HR, thus introducing errors in HR estimation. More sophisticated denoising techniques are thus required to obtain clean heart sounds under motion. The discrete wavelet transform (DWT) 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 by the MAs to some extent, due to the underlying complexity of the artefacts, and the closely overlapping frequency ranges of the artifacts and the heart sounds. Therefore, we propose a frequency spectrum searching algorithm to estimate the HR from the denoised signal to account for the remaining MAs. Instead of searching the FFT peaks over the full frequency range of the denoised audio, we only search the HR peaks in a small frequency range corresponding to the range of allowable human HRs and the HR in the previous window. This guarantees that peaks in HR-unrelated frequency ranges are not taken as HR and the HR 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 HR. Thus, we acknowledge that a more sophisticated approach to the problem, specifically to addressing signal denoising, is required.

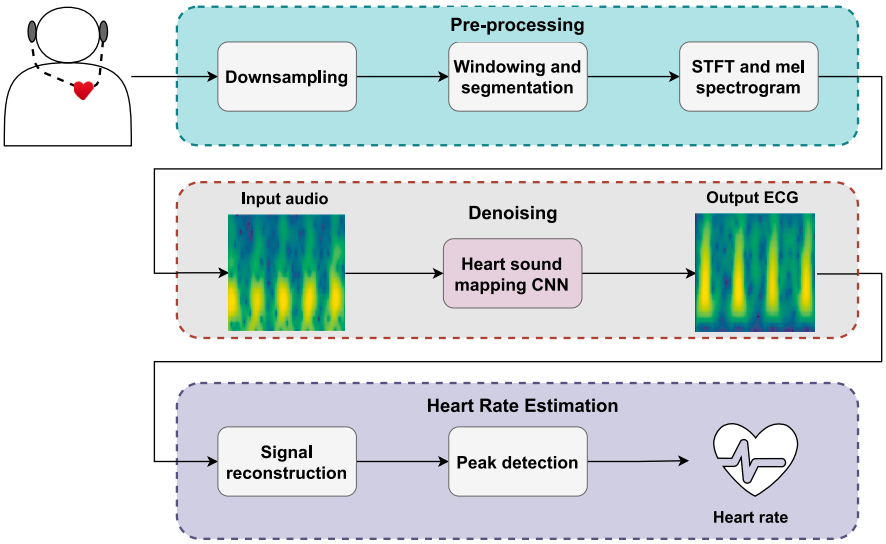


Fig. 4. hEARt system flowchart.

3.2. Overview of the deep learning-based pipeline

An overview of hEARt, our designed motion-resilient HR monitoring system, is given in Fig. 4. Audio signals captured inside the occluded ear canal are used for HR estimation, which is performed in three stages: pre-processing, MA elimination and HR estimation. Pre-processing aims at removing the frequency components unrelated to heart sounds. For MA elimination, we proposed a CNN-based network to map spectrograms of noisy heart sounds signal to spectrograms of the corresponding ECG (a clean heart signal) during the training phase, thus producing an output *synthesized ECG*. Our problem is thus framed as a denoising problem, but also as a synthesis problem. We adopted a U-Net encoder–decoder architecture for denoising since audio (and specifically heart sounds) is commonly represented in image form as spectrograms [23–25]. Initially developed for biomedical image segmentation, U-Net shows great potential in image denoising and super resolution [23,26]. 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 [27], while complex network structures [28,29] could easily suffer from overfitting. Finally, HR is estimated using peak detection on the clean signals.

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 and segment the audio into 2 s windows, each with a 1.5 s overlap with the previous window. 2 s windows were selected to ensure the presence of multiple heart beats (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 HR calculation, including music and ambient noise. Additionally, due to 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 2.2, MAs and other interfering signals lie overlapping frequency ranges with heart sounds, therefore requiring additional processing.

We process the GT 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 emphasize the peaks in the ECG (the QRS complex) while attenuating the P and T waves (as seen in Fig. 2). Since we are only interested in capturing the beats and the inter-beat timing (for measuring HR, and in future, heart rate variability), only the QRS complex is of interest.

3.4. Motion artefact elimination

The MA elimination subsystem takes as input pre-processed audio signals, and outputs cleaned heart signals. To do so, it uses the GT ECG signal to supervise the denoising of the heart signals.

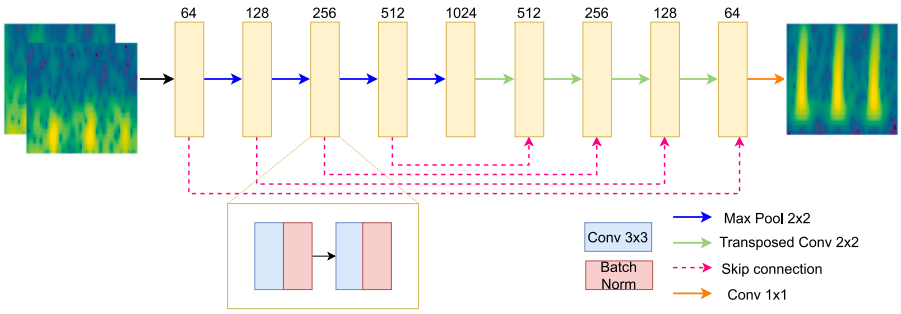


Fig. 5. U-Net autoencoder architecture.

3.4.1. Spectrogram generation

We compute log-mel spectrograms of the windowed audio and ECG signals using short-time Fourier transform (STFT), with a window size of 256 samples and hop length of 32 samples. 1024 FFT bins are used with zero padding and a Hann window. 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 sounds 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 $64 \times 64 \times 2$ input. The output is a single channel ECG spectrogram. The spectrograms are normalized between 0 to 1, to aid network training. Normalization is carried out by dividing by a constant value, to maintain the difference in the signal amplitude for different activities.

3.4.2. Network structure

Fig. 5 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 normalization 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 normalization layers as in the encoder. In the final layer, a 1×1 convolution is used to map the feature maps into a single 64×64 output image.

3.4.3. 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 [30], where log-mel spectrograms of heart sounds are used 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 initialization weights for the CNN, which is fine-tuned using our data. This helps leverage and transfer the knowledge learnt about heart sounds using PASCAL, as well as avoiding overfitting on a small dataset.

3.4.4. 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 trained on the other 19 users. The model is trained empirically for 100 epochs using the Adam optimizer 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.

The system uses the structural similarity index measure (SSIM) as the loss function [31]. SSIM is a measure of image similarity, which takes into account local structure of the image. SSIM has been successfully used for image reconstruction and image denoising [31], making it suitable for our task. SSIM loss ($L_{SSIM}(P)$) is defined in Eq. (1), where the SSIM ($SSIM(p)$) of two windows x and y is defined in Eq. (2). In Eq. (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 [31].

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

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

3.4.5. Signal reconstruction

We convert the reconstructed clean spectrograms to time-domain waveforms for HR estimation. The Griffin–Lim algorithm [32] 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.

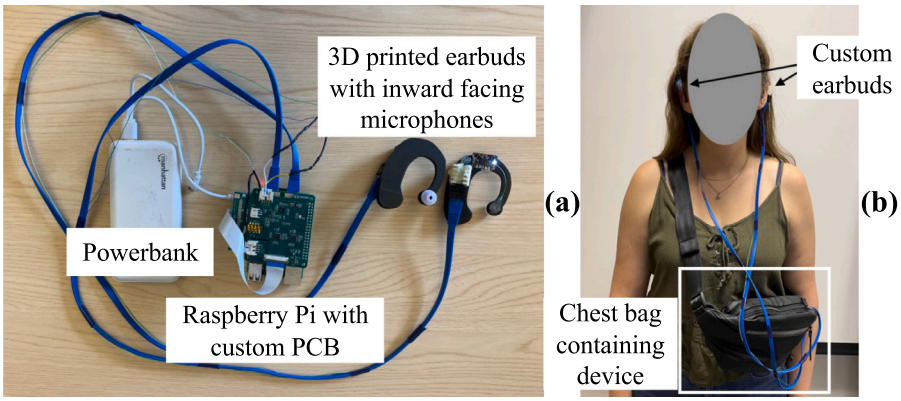


Fig. 6. (a) Custom hardware prototype and (b) participant wearing the device.

3.5. Heart rate estimation

HR estimation is performed on a 10 s long window, where each window has a 6s overlap with the previous window [33]. 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 HR for the window. Finally, a moving average window of 5 samples is used to remove outliers from the predictions.

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.

4.1. Prototyping

While in-ear microphones have been successfully integrated into existing commercial earbuds, such as the AirPods Pro, 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 (Fig. 6). 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 [34] 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. On the left earbud, we embedded a speaker behind the microphone to enable audio playback.

To capture signals from the microphones, we designed a PCB which interfaces with an audio codec [35] onto a Raspberry Pi 4B. The PCB contains a MCP6004 non-inverting operational amplifier with adjustable gain controlled using potentiometers for additional flexibility. The microphone signals are thus amplified before being sampled by the audio codec onto the Raspberry Pi. The system is shown in Fig. 6. To make the system portable, we placed the Raspberry Pi and PCB into a chest-worn bag and power it with a portable power bank. This ensures that the prototype does 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 movement of the wires.

4.2. Data collection

We used a Zephyr BioHarness 3.0 chest strap [36] to measure the GT heart signal. Specifically, we extracted the raw ECG from the Zephyr BioHarness and use it as both the clean heart signal for the CNN and to calculate the GT HR. 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 HR estimation is performed in 10 s windows, this delay is negligible.

We invited 15 participants (8 males and 7 females) for data collection.[1] We ran an extensive data collection protocol consisting of stationary and motion scenarios, as shown in Table 1 amounting to a total of 41 min of data per participant. These activities were selected as they encompass typical scenarios when a person uses earbuds and requires HR monitoring. Our data collection involved

[1] The experiment has been approved by the Ethics Committee of the institution.

Table 1
Data collection protocol.

Activity	Duration (minutes)	Mean HR	M in HR	Max HR
Stationary				
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

controlled activities and also in-the-wild scenarios to ensure the applicability of our methods to real-world use. 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 2 km/h and 5 km/h. For running, between 5 km/h and 8 km/h. The average sound level in the room was 32 dB while stationary and 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 cool down HR. While sitting, users were asked to move their heads three times to capture head motions to ensure that our stationary HR estimation is accurate even under the presence of non-full body motions. While meditating, participants were given a video to follow with controlled inhalation and exhalation durations to try slow the HR.

Table 1 also provides the mean, minimum and maximum HR for each activity. It is evident that the distribution of HR varies per activity, with running having the largest range of HR. These activities also clearly represent a large range of HR, with HR of up to 144 BPM even while stationary. This dataset is thus well suited to test the ability of the system to predict HR under a wide range of conditions.

5. Performance evaluation

5.1. Metrics

We evaluated the performance of our system according to the following metrics [37]:

(i) Mean Absolute Error (**MAE**): the average absolute error between the GT HR (BPM_{true}) and the calculated HR (BPM_{calc}) for each window ($i, i \in [1, N]$) (Eq. (3)).

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |BPM_{calc}(i) - BPM_{true}(i)| \tag{3}$$

(ii) Mean Average Percentage Error (**MAPE**): the average percentage error over each window (Eq. (4)).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \frac{|BPM_{calc}(i) - BPM_{true}(i)|}{BPM_{true}(i)} * 100 \tag{4}$$

(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. HR from the GT). 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 [38]. BA plots are used clinically to assess the level of agreement between two measurement methods [38]. In this work, we compare the calculated HR to the GT HR for each 10 s window.

5.2. Baseline comparison

Table 2 shows the performance comparison between the proposed DL-based hEARt system and two signal processing approaches - (1) the proposed signal processing method (referred to as SP) leverages the DWT for signal denoising and extracts HR from the frequency spectrum of the denoised signals. (2) we additionally compare our methods to the baseline developed by Martin and Voix [8] (referred to as baseline), which uses Hilbert transforms and peak detection for HR estimation in the time domain, under stationary conditions. To perform this comparison, we group our activities into three scenarios based on the level of full body activity: (1) stationary (comprising of sitting, standing, lying down, listening to music, meditation, working and cooling down after exercise); (2) walking; (3) running. Our proposed SP approach outperforms the baseline for all activities. This demonstrates that the baseline algorithm designed for stationary is unable to generalize to motion conditions, and an additional denoising module is required. Additionally, it indicates that the motion artefacts introduced by in-the-wild stationary activities greatly deteriorate

Table 2
Comparison between hEARt, the two baselines and in-ear PPG in terms of MAPE (%).

Activity	hEARt	Signal processing	Baseline [8]	In-ear PPG [7]
Stationary	**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

Table 3
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

performance of HR estimation, meaning that even while stationary more sophisticated techniques are needed. Comparing the SP approach with hEARt, we observe that hEARt significantly outperforms SP for each of the activities, showing that the DL based technique is better at generalizing to the differences in the data, and to motion artefacts, than the SP approach. From our previous work [39], when considering only sitting, performance was comparable using the SP and DL approaches. This is because sitting is a highly controlled activity with minimal movement. With the addition of extra, less controlled, stationary activities, the SP approach fails even in the stationary case due to the increase in motion artefacts and thus signal noise.

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

Table 2 also compares the performance of hEARt with that of in-ear PPG (as studied by Ferlini et al. [7]). It is evident from the table that (i) although PPG is the gold-standard for HR measurement, full-body motion causes significant degradation in HR 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 HR measurement through the ear.

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. The average MAE of the collection of stationary activities is 1.88 ± 2.89 BPM. From the table, it is evident that, as expected, full body motion degrades HR estimation ability with errors increasing with increasing intensity of motion. By examining the stationary activities, we also see the impact of motion artefacts on estimation error. Specifically, working in the wild has the largest error of the stationary activities, which is intuitive since any in-the-wild scenario will inherently contain more movement (thus more artefacts in the signal) than a controlled scenario. 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 [14].

One notable result is the excellent performance while playing music through the speaker on the earable device. Even in the presence of music, HR 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.

5.4. Individual HR estimation

Next, we evaluate our approach under different activities for all participants. First, we provide some insights on the population statistics. Fig. 7(a) reports a heatmap of the MAPE of the audio-extracted HR for every user across the activities. Lighter colours 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: **(i)** errors for motion conditions are higher than stationary. **(ii)** our system generalizes well to the different activities with mostly consistent errors amongst participants for the activities. **(iii)** 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 generalize to different users and that with high quality data, good HR estimation can be achieved.

To further assess the performance of hEARt on individuals, we provide a boxplot of the MAPE of HR estimation overall activities for each participant in Fig. 7(b). It is evident that there is little variation in the median error of each participant, with the median

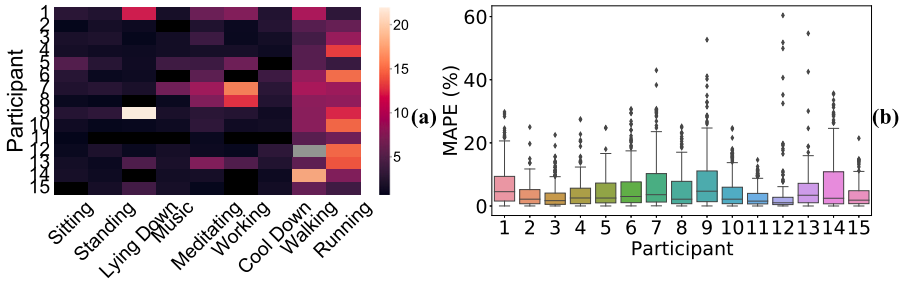


Fig. 7. (a) MAPE heatmap per participant and (b) boxplot of MAPE per participant.

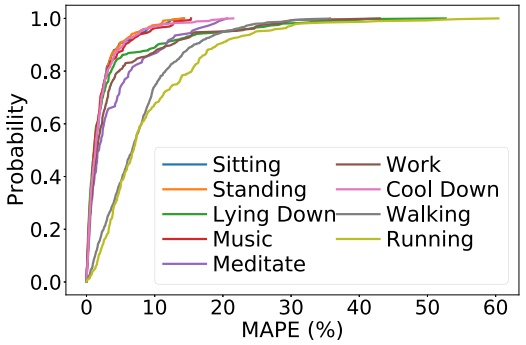


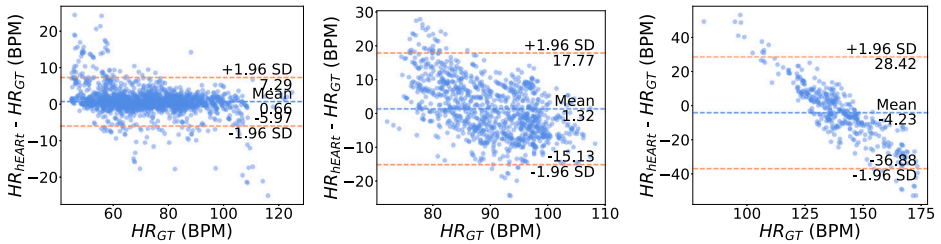
Fig. 8. Empirical CDF of the MAPE of heart rate estimation over the different activities.

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 the higher median error. Participant 9 has a very large error for lying down which results in the larger median error. 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. Likewise, participant 9 has a large number of outliers due to the very large error while lying down. This error is due to weak signals when lying due to the displacement of the earbuds. However, overall, the results show that the system is able to generalize to different users over different activities.

To further understand the extent to which the various activities impact hEARt, for each of them we report the empirical cumulative distribution function (ECDF) of the error (Fig. 8). These results align with those in Fig. 7, 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 HR monitoring offers a promising modality for continuous HR sensing in presence of motion.

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 HR calculated with hEARt and that obtained from the GT chest strap) for the three scenarios based on activity level (stationary, walking, running) in Fig. 9. Specifically, Fig. 9(a) reports the agreement while stationary. It is clear that the bias between the two measurements is minimal, with very low mean (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 (Fig. 9(b) and Fig. 9(c) respectively), wider limits of agreement are present, representing a greater standard deviation in the HR estimation. For walking, the positive mean error shows a slight overall tendency for overestimation. Additionally, there is a slight trend where lower HR 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 HR and a tendency to slightly underestimate HR. This can also be traced down to the imbalance of our dataset, where lower HR values are predominant, even with the longer durations of the walking and running activities compared to [39]. Especially in the running case, this underestimation of higher HR is observed when the frequency of the running overlaps with the HR values. The MA-induced spikes trigger a harsher response by hEARt which tries to remove the noisy peaks, thus leading to an underestimation of higher HR. Nonetheless, our approach still performs well for all activities.



(a) Stationary (b) Walking (c) Running

Fig. 9. Modified Bland–Altman plot of heart rate extraction.

Table 4
Performance of hEARt in outdoor settings.

Activity	MAPE (%)	SAPE (%)
Stationary	**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**

5.6. Outdoor performance

To ensure the applicability of our system to real-world scenarios and diverse ambient noise levels, we assessed performance on 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 in the vicinity of an active building construction site, where the ambient sound level averaged 53 dB. For the walking and running assessments, participants were given the freedom to select their preferred pace and move freely within the designated area. This setup aimed to replicate real-life situations and variations in noise levels, making our evaluation more comprehensive and reliable.

Table 4 provides the results for the outdoor study. By comparing Table 2, 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 HR 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.

5.7. 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 HR prediction for one user in this study is given in Fig. 10. From the figure, it is clear that the system is able to accurately predict HR even in uncontrolled environments as the trends of the two lines very closely match one another. We see slight underestimations of the HR while walking, however the overall estimations are accurate.

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 2, 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 generalizable to different conditions and to different activities. Thus this study acts as a proof of concept of the in-the-wild feasibility of hEARt.

5.8. Power and latency measurements

For a comprehensive system analysis, we evaluated the power consumption and latency of our implemented system 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 5 provides a breakdown of the operation times for the various system

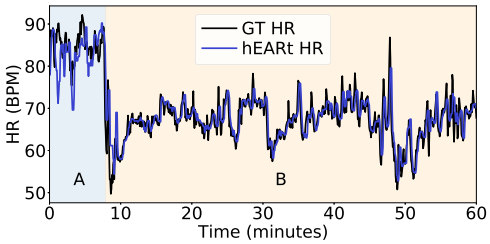


Fig. 10. Longitudinal heart rate tracking. Coloured boxes indicate the different activities. A: Walking. B: Working in the wild.

Table 5
Latency.

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

Table 6
Power consumption.

Operation	Power (W)
RasPi (Baseline)	2.870
RasPi+Mic	3.032
Full system	3.361

components. Signal denoising was performed on a 2 s window and HR extraction on a 10 s window, as detailed in Section 3. Processing a 10 s window takes the system 347 ms, implying that a new HR can be predicted by the system every 4s (due to the 6s second overlap). This latency parameter is adjustable based on the desired overlap ratio.

The system power consumption is given in Table 6. Overall, the system (including microphone sampling, denoising and HR 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 4s. Thus, the average energy consumed per second is $(3032-2870)\ \text{mW} \times 1\ \text{s} + (3361-3032)\ \text{mW} \times 83.06\ \text{ms}/4 = 699\ \text{mJ}$. Although this energy consumption might appear substantial, it is worth noting that our system was implemented on a power-intensive Raspberry Pi without optimizing 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, optimizations and quantization were not applied. The model can thus be further optimized to reduce energy consumption and latency, thus lowering energy expenditure. These optimizations will be required before hEARt can feasibly be used on a commercial earbud.

6. Discussion

While we acknowledge the merits of PPG-based HR monitoring and are aware of the wealth of information PPG carries, there is great value in showing the potential of a lesser explored modality: in-ear microphones. In-ear microphones offer substantial advantages over PPG, including their price tag and prevalence in high-end earbuds and hearing aids, due to their importance in adaptive noise cancellation. Microphones are also relatively power efficient sensors [15], requiring less current than PPG (especially when used with high intensity configurations to increase SNR) [16]. Concretely, the microphone used in our prototype [15] has a current draw of 0.12 mA, more than 10 times less than that of a state-of-the-art wearable dedicated PPG module, the MAXM86161, which draws 1.62 mA [16].

In this work, we use the same approach used previously in [39], but with a much larger dataset consisting of more activities and longer activity duration. The result of this being that a larger range of HR are represented in the dataset. This has improved the ability of the model to generalize to different HR, as shown by the result improvement between this work and [39]. We thus expect that with the collection of further data, the results can be further improved.

In the remainder of this section we reason over some shortcomings of our work, and potential solutions. First, we are aware of the limitations that come with a simple, cheap prototype like ours. For instance, a few pieces of the collected data were corrupted, or had worse signal quality, as the subjects were unable to achieve a good fit with the earbuds. This indicates that proper sealing of the ear canal is critical. Given that people have different shaped and sized ear canals, it is necessary to select the optimal ear tip size for each individual to improve performance, using an automated method of checking the fit of the earbuds and the seal as

done in [10] or as implemented by Apple for the AirPod Pro.[2] The data quality during running was also worsened by the wires on the earbuds which move during vigorous activity thus dislodging or loosening the earbuds. Using a wireless prototype would thus improve earbud fit and resulting system performance. Interestingly, fit and positioning issues have also been reported for in-ear PPG [7]. Though, contrary to PPG where sensor misplacement can be hard to identify and may lead to artefacts, poor fit is obvious with in-ear microphones [10]. Nonetheless, our work shows the viability of using in-ear microphones for the detection of HR, even with a far-from-optimal academic prototype.

Given that earbuds are mainly used for audio delivery, one concern is whether music playback will affect performance. However, as shown in Table 3, listening to music does not affect stationary HR estimation, with very low errors being obtained even with music playback. However, for future work, the ability of the system to estimate HR under a combination of conditions must be assessed, e.g. listening to music while walking and running.

Finally, despite the good performance, more strategies can be utilized for further improvement. Firstly, we expect that fine tuning a model for each activity will improve activity level performance. We also aim to investigate the use of a LSTM-based model to better model dependencies between adjacent heart sounds. Additionally, collecting data from more subjects encompassing a wider range of HR, specifically introducing longer durations of walking and running, will improve the ability of the model to further generalize to higher HR. Finally, we aim to implement automated seal quality detection which detects if the seal quality is too poor to do reliable HR extraction. In this case, the user can be notified to reposition their earbuds to improve signal quality. This will ensure that in situations such as lying down, where earbuds have a tendency to be dislodged, accurate monitoring can still be achieved.

7. Related work

7.1. Earables

Earables have attracted tremendous attention for human sensing applications, especially for health and wellbeing monitoring [5]. Literature has investigated earables for blood flow and oxygen consumption [40], dietary monitoring and swallow detection [41], blood pressure monitoring [42], step counting [10], heart and respiratory rate tracking [8], user identification and gesture recognition [43], etc. Bui et al. [42] proposed a device to measure blood pressure from the artery in the ear canal. Amft et al. [41] developed a monitoring system that classifies different kinds of food by analysing chewing sounds. With respect to motion tracking, facial expression tracking shows great potential, due to the physical deformations generated by facial muscle movements [44,45]. An acoustic-based in-ear system using the in ear microphone for step counting, activity recognition, and hand-to-face gesture interaction was also investigated [10]. Respiratory rate measuring and biological analysis are also two vital application fields for earables [46–48]. A paradigm named HeadFi was proposed to turn the drivers inside existing headphones into a sensor, with its potential validated in four applications [43]. Using the HeadFi system, the authors perform HR monitoring in the stationary case and with the addition of body movement caused by taking the headphones on and off. However, they did not study HR monitoring in the presence of full-body motion such as running and walking.

7.2. Heart rate monitoring

HR is generally measured using electroencephalogram (EEG), ECG or PPG sensors. However, EEG has limited applications out-of-the-clinic and ECG requires a chest strap, making it inconvenient. Some smartwatches can capture ECGs, but require the user to close the circuit with their fingers. PPG is the standard for HR monitoring in wearables. However, it is highly susceptible to MAs caused by physical activity or body motion [37]. [2] showed that amongst consumer and research grade wrist-worn wearables, the error of HR estimation was 30% higher during activity than at rest. A particular problem with PPG is the signal crossover effect where the PPG sensors lock onto a periodic signal from motion (such as walking or running), which is mistaken as the heart signal [2,3] causing measurement errors. Recently, [7] reported a 27.14% and 29.84% error of PPG sensors in earables for walking and running respectively, quantitatively demonstrating the challenges of PPG in HR estimation under motion. Acoustic sensors have also been studied for HR measurements. Chen et al. [49] estimated HR from a small acoustic sensor placed at the neck, and [50] measured in-ear pulse wave velocity using heart signals as reference. [8] examines both heart and breathing rates using microphones placed in the ear canal while stationary. Artefacts were found due to minor movement of the subject's body even though all recordings are collected with subjects remaining stationary. [51] introduces a earphone that is equipped with an in-ear microphone to measure HR and an IMU to measure activity level. However, the impact of activity-induced vibrations on the in-ear heart sounds is not investigated. These findings imply the challenges in HR measurement from earables under motion. In our previous work [52], we have presented an approach that aims to tackle these challenges and offer a solution to measuring HR in realistic settings: hEARt, an in-ear audio based system for HR monitoring under four scenarios (sitting still, walking, running and speaking). In this work, we further prove the real-world applicability of the system. Namely, we developed a new earable prototype with improved signal quality and we collected an extensive new dataset from 15 subjects with nine activities per subject and much longer evaluation periods than in [52]. Our new dataset also encompasses a much larger range of heart rates, enabling a richer evaluation system performance. Additionally, our new dataset contains audio collected under different ambient noise levels and while listening to music. We have thus presented a more complete system evaluation, and have improved upon our previous results through having more, better quality data.

[2] https://support.apple.com/en-us/HT210633

8. Conclusion

We proposed an approach for accurate HR estimation using audio signals collected in the ear canal, under motion artefacts caused by daily activities (i.e. walking, and running). Specifically, leveraging deep learning, we eliminate the interference of motion artefacts and recreate clean heart signals, from which we are able to determine HR. We designed a prototype and collected data from real subjects to evaluate the system. Experimental results demonstrate that our approach achieves mean absolute errors of 1.88 ± 2.89 BPM, 6.83 ± 5.05 BPM, and 13.19 ± 11.37 BPM for stationary, walking, and running, respectively, opening the door to new non-invasive and affordable HR monitoring with useable performance for daily activities. We also discussed some potential strategies to further improve the performance in the future.

CRediT authorship contribution statement

Kayla-Jade Butkow: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ting Dang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Andrea Ferlini:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Dong Ma:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Yang Liu:** Data curation, Investigation. **Cecilia Mascolo:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported by ERC through Project 833296 (EAR), U.K. EPSRC Centre for Doctoral Training in Sensor Technologies for a Healthy and Sustainable Future (EP/S023046/1), the Cambridge Trust, Nokia Bell Labs through a donation and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Grant ID: 21-SIS-SMU-036, 001124-00001).

References

[1] S. Michael, K.S. Graham, G.M. Davis, Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals-A review, Front. Physiol. 8 (2017) 301, http://dx.doi.org/10.3389/fphys.2017.00301.
[2] B. Bent, B.A. Goldstein, W.A. Kibbe, J.P. Dunn, Investigating sources of inaccuracy in wearable optical heart rate sensors, npj Digit. Med. 3 (1) (2020) 18, http://dx.doi.org/10.1038/s41746-020-0226-6.
[3] J.W. Navalta, J. Montes, N.G. Bodell, R.W. Salatto, J.W. Manning, M. DeBeliso, Concurrent heart rate validity of wearable technology devices during trail running, Plos One 15 (8) (2020).
[4] J. Ahn, H.-K. Ra, H.J. Yoon, S.H. Son, J. Ko, On-device filter design for self-identifying inaccurate heart rate readings on wrist-worn PPG sensors, IEEE Access 8 (2020) 184774–184784.
[5] F. Kawsar, C. Min, A. Mathur, A. Montanari, Earables for personal-scale behavior analytics, IEEE Pervasive Comput. 17 (3) (2018) 83–89, http://dx.doi.org/10.1109/MPRV.2018.03367740.
[6] V. Goverdovsky, W. Von Rosenberg, T. Nakamura, D. Looney, D.J. Sharp, C. Papavassiliou, M.J. Morrell, D.P. Mandic, Hearables: Multimodal physiological in-ear sensing, Sci. Rep. 7 (1) (2017) 1–10.
[7] A. Ferlini, A. Montanari, C. Min, H. Li, U. Sassi, F. Kawsar, In-ear PPG for vital signs, IEEE Pervasive Comput. 21 (1) (2022) 65–74, http://dx.doi.org/10.1109/MPRV.2021.3121171.
[8] A. Martin, J. Voix, In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring, IEEE Trans. Biomed. Eng. 65 (6) (2018) 1256–1263, http://dx.doi.org/10.1109/TBME.2017.2720463.
[9] M.A. Stone, A.M. Paul, P. Axon, B.C. Moore, A technique for estimating the occlusion effect for frequencies below 125 Hz, Ear Hear. 35 (1) (2014) 49–55, http://dx.doi.org/10.1097/AUD.0b013e31829f2672.
[10] D. Ma, A. Ferlini, 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, Association for Computing Machinery, New York, NY, USA, 2021, pp. 175–187, http://dx.doi.org/10.1145/3458864.3467680.
[11] M. Murray, G. Spurr, S. Sepic, G. Gardner, L. Mollinger, Treadmill vs. floor walking: kinematics, electromyogram, and heart rate, J. Appl. Physiol. 59 (1) (1985) 87–91.
[12] S. Passler, N. Müller, V. Senner, In-ear pulse rate measurement: A valid alternative to heart rate derived from electrocardiography? Sensors 19 (17) (2019) 3641.
[13] J.A. Patterson, D.C. McIlwraith, 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.
[14] Consumer Technology Association, Physical activity monitoring for heart rate ANSI/CTA-2065, 2018.
[15] Zero-height SiSonic microphone with extended low frequency performance, (SPU1410LR5H-QB) 2013, Rev. D.
[16] Single-supply integrated optical module for HR and SpO2 measurement, (MAXM86161) 2019, Rev. 0.
[17] J. Tonndorf, A new concept of bone conduction, Arch. Otolaryngol. 87 (6) (1968) 595–600.

[18] A. Ferlini, D. Ma, R. Harle, 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.

[19] M. Ali, P. 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.

[20] M.N. Ali, E.-S.A. El-Dahshan, A.H. Yahia, Denoising of heart sound signals using discrete wavelet transform, Circuits Systems Signal Process. 36 (11) (2017) 4482–4497.

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

[22] 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.

[23] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, Springer International Publishing, Cham, 2015, pp. 234–241.

[24] M.T. Nguyen, W.W. Lin, J.H. Huang, Heart sound classification using deep learning techniques based on log-mel spectrogram, Circuits Systems Signal Process. (2022) http://dx.doi.org/10.1007/s00034-022-02124-1.

[25] F. Demir, A. Şengür, V. Bajaj, K. Polat, Towards the classification of heart sounds based on convolutional deep neural network, Health Inf. Sci. Syst. 7 (1) (2019) 16, http://dx.doi.org/10.1007/s13755-019-0078-0.

[26] W. Xu, X. Deng, S. Guo, J. Chen, L. Sun, X. Zheng, Y. Xiong, Y. Shen, X. Wang, High-resolution U-net: preserving image details for cultivated land extraction, Sensors (Basel, Switzerland) 20 (15) (2020) 4064, http://dx.doi.org/10.3390/s20154064.

[27] M.Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, V.K. Asari, Recurrent residual convolutional neural network based on U-net (R2U-net) for medical image segmentation, 2018, http://dx.doi.org/10.48550/ARXIV.1802.06955, URL https://arxiv.org/abs/1802.06955.

[28] J. Yoon, D. Jarrett, M. van der Schaar, Time-series generative adversarial networks, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2019.

[29] Q. Man, Y.-I. Cho, S.-G. Jang, H.-J. Lee, Transformer-based GAN for new hairstyle generative networks, Electronics 11 (13) (2022) http://dx.doi.org/10.3390/electronics11132106.

[30] P. Bentley, G. Nordehn, M. Coimbra, S. Mannor, The PASCAL classifying heart sounds challenge 2011 (CHSC2011) results, http://www.peterjbentley.com/heartchallenge/index.html.

[31] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for neural networks for image processing, (arXiv:1511.08861) 2018, arXiv:1511.08861.

[32] N. Perraudin, P. Balazs, P.L. Sø ndergaard, A fast griffin-lim algorithm, in: 2013 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics, 2013, pp. 1–4, http://dx.doi.org/10.1109/WASPAA.2013.6701851.

[33] Q. Li, R.G. Mark, G.D. Clifford, Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter, Physiol. Meas. 29 (1) (2008) 15–32, http://dx.doi.org/10.1088/0967-3334/29/1/002.

[34] SPU1410lr5h-QB, Digi-Key Electron. (2023) https://www.digikey.co.uk/en/products/detail/knowles/SPU1410LR5H-QB/3621629.

[35] Respeaker 6-mic circular array kit for raspberry pi | seeed studio wiki, 2023, https://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/.

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

[37] S. Ismail, U. Akram, I. Siddiqi, Heart rate tracking in photoplethysmography signals affected by motion artifacts: A review, in: Eurasip Journal on Advances in Signal Processing, vol. 2021, (1) Springer Science and Business Media Deutschland GmbH, 2021, p. 5, http://dx.doi.org/10.1186/s13634-020-00714-2.

[38] D. Giavarina, Understanding bland altman analysis, Biochem. Med. 25 (2) (2015) 141–151, http://dx.doi.org/10.11613/BM.2015.015.

[39] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, C. Mascolo, Heart: Motion-resilient heart rate monitoring with in-ear microphones, in: 2023 IEEE International Conference on Pervasive Computing and Communications, (PerCom), 2023, pp. 200–209, http://dx.doi.org/10.1109/PERCOM56429.2023.10099317.

[40] S.F. LeBoeuf, M.E. Aumer, W.E. Kraus, J.L. Johnson, B. Duscha, Earbud-based sensor for the assessment of energy expenditure, heart rate, and VO2max, Med. Sci. Sports Exercise 46 (5) (2014) 1046–1052, http://dx.doi.org/10.1249/MSS.0000000000000183.

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

[42] 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, 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, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–17, http://dx.doi.org/10.1145/3300061.3345454.

[43] X. Fan, L. Shangguan, S. Rupavatharam, Y. Zhang, J. Xiong, Y. Ma, 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.

[44] D.J. Matthies, B.A. Strecker, B. Urban, Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 1911–1922.

[45] T. Ando, Y. Kubo, B. Shizuki, S. Takahashi, Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals, in: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, 2017, pp. 679–689.

[46] T. Röddiger, D. Wolffram, D. Laubenstein, M. Budde, M. Beigl, Towards respiration rate monitoring using an in-ear headphone inertial measurement unit, in: Proceedings of the 1st International Workshop on Earable Computing, ACM, London United Kingdom, 2019, pp. 48–53, http://dx.doi.org/10.1145/3345615.3361130.

[47] G. Pressler, J. Mansfield, H. Pasterkamp, G. Wodicka, Detection of respiratory sounds at the external ear, IEEE Trans. Biomed. Eng. 51 (12) (2004) 2089–2096, http://dx.doi.org/10.1109/TBME.2004.836525, URL http://ieeexplore.ieee.org/document/1360027/.

[48] K. Jafarian, K. Hassani, D.J. Doyle, M.N. Lahiji, O.M. Moghaddam, A. Saket, M. Majidi, F. Izadi, Color spectrographic respiratory monitoring from the external ear canal, Clin. Sci. 132 (24) (2018) 2599–2607, http://dx.doi.org/10.1042/CS20180748.

[49] G. Chen, S.A. Imtiaz, E. Aguilar-Pelaez, E. Rodriguez-Villegas, Algorithm for heart rate extraction in a novel wearable acoustic sensor, Healthc. Technol. Lett. 2 (1) (2015) 28–33.

[50] R. Kusche, P. Klimach, A. Malhotra, S. Kaufmann, M. Ryschka, An in-ear pulse wave velocity measurement system using heart sounds as time reference, Curr. Dir. Biomed. Eng. 1 (1) (2015) 366–370.

[51] S. Nirjon, R.F. Dickerson, Q. Li, P. Asare, J.A. Stankovic, D. Hong, B. Zhang, X. Jiang, G. Shen, 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.

[52] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, C. Mascolo, Heart: Motion-resilient heart rate monitoring with in-ear microphones, in: 2023 IEEE International Conference on Pervasive Computing and Communications (PerCom), 2023, pp. 200–209, http://dx.doi.org/10.1109/PERCOM56429.2023.10099317.