3-2023

Heart: Motion-resilient heart rate monitoring with in-ear microphones

Kayla-Jade BUTKOW

Ting DANG

Andrea FERLINI

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

MASCOLO

hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones

Kayla-Jade Butkow*, Ting Dang*, Andrea Ferlini*, Dong Ma*† and Cecilia Mascolo*
*University of Cambridge, Cambridge, UK †Singapore Management University, Singapore
{kjb85, td464, af679}@cam.ac.uk, dongma@smu.edu.sg, cm542@cam.ac.uk

2023 IEEE International Conference on Pervasive Computing and Communications (PerCom) | 978-1-6654-5378-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/PERCOM56429.2023.10099317

***Abstract*—With the soaring adoption of in-ear wearables, the research community has started investigating suitable in-ear heart rate (HR) detection systems. HR is a key physiological marker of cardiovascular health and physical fitness. Continuous and reliable HR monitoring with wearable devices has therefore gained increasing attention in recent years. Existing HR 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* HR monitoring. We first collected HR-induced sounds in the ear canal leveraging an in-ear microphone under stationary and three different activities (i.e., walking, running, and speaking). Then, we devised a novel deep learning based motion artefact (MA) mitigation framework to denoise the in-ear audio signals, followed by an HR estimation algorithm to extract HR. With data collected from 20 subjects over four activities, we demonstrate that hEARt, our end-to-end approach, achieves a mean absolute error (MAE) of 3.02 $\pm$ 2.97 BPM, 8.12 $\pm$ 6.74 BPM, 11.23 $\pm$ 9.20 BPM and 9.39 $\pm$ 6.97 BPM for stationary, walking, running and speaking, respectively, opening the door to a new non-invasive and affordable HR monitoring with usable performance for daily activities. Not only does hEARt outperform previous in-ear HR monitoring work, but it outperforms reported in-ear PPG performance.**

***Index Terms*—earable, heart rate, motion artefact, in-ear audio**

I. 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 (HRV), the change in time between successive beats, 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 [1]. Thus measuring HR under motion is critical for monitoring human health and wellbeing.

Electrocardiographic (ECG) telemetry monitoring is the standard for HR and HRV 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 smartwatches include a single-lead ECG, however they require the user 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

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).

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 (e.g., walking, running and speaking). 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 three challenges. First, the strength of heartbeats is much weaker than the foot strikes, so heartbeat signals are buried in the walking signals. 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. Third, due to the proximity of the ear to the human vocal system, human speech and its associated jaw movements can overwhelm the heartbeat-induced sound.

To address these challenges, we propose a processing pipeline for accurate HR detection in the presence of MAs, namely, walking, running and speaking. Different from previous audio-based HR estimation works [6], [8], [12], [13], we also validate the functioning of our technique in the presence of speech, showing how the proposed approach can successfully deal with speaking activities. With data collected from 20 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] (65% and 67% 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 common human activities. **(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 (HS) 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 HS dataset and fine-tunes it using our data to effectively capture HS 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 HS captured by earables using ECG signals. **(iii)** We built an earbud prototype with good signal-to-noise ratio (SNR) and collected data from 20 subjects. Results show that we can achieve mean absolute errors of 3.02 $\pm$ 2.97 BPM, 8.12 $\pm$ 6.74 BPM, 11.23 $\pm$ 9.20 BPM and 9.39 $\pm$ 6.97 BPM for stationary, walking, running and speaking, respectively, demonstrating the effectiveness of the proposed approach in combating MAs.

II. Primer

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

A. In-ear Heart Sound Acquisition

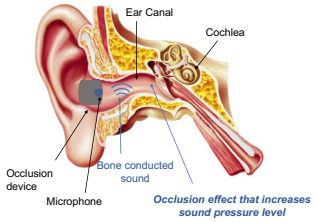


Fig. 1: The occlusion effect and the anatomy of the ear.

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 Figure 1, is known as the occlusion effect. Since bone conveys low-frequency sounds [17], the bone-conducted HS are amplified in the occluded ear canal [8]. HS can thus be detected using a microphone placed inside the occluded ear canal. An example showing the HS captured by the internal microphone is shown in Figure 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 Figure 2.

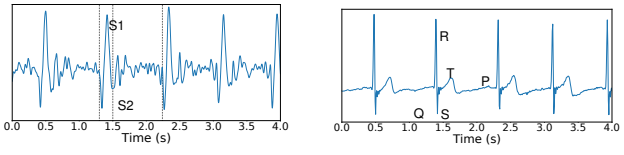
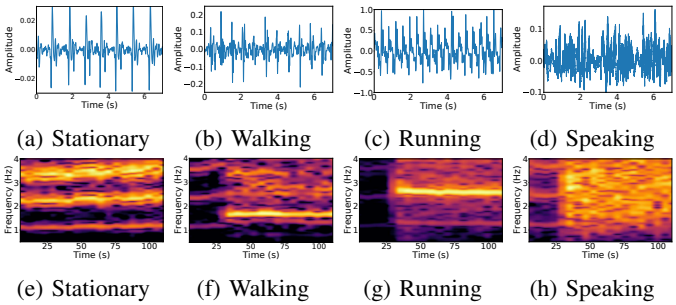


Fig. 2: The (left) sound signal captured by the internal microphone and the (right) corresponding ECG signal.

B. 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



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

(e) Stationary (f) Walking (g) Running (h) Speaking

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

and vibrations inside the body [10], [18]. Figures 3a to 3d illustrates the recorded audio signals from the in-ear microphone while stationary, walking, running and speaking within a seven second window. Figures 3e to 3h 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 Figure 3a, with frequency lying around 1.2Hz and its 1st and 2nd harmonics clearly observable in Figure 3e. Contrastingly, it is completely overwhelmed by the amplified step sounds in Figure 3b (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 1.7Hz (the cadence) in Figure 3f. Though the HR and its harmonics are still observable, it is difficult to estimate HR directly from the raw corrupted audio signals. Furthermore, it is evident that the periods of HS and walking are similar, resulting in an overlap in the frequency domain, making it challenging to split the HS and walking signals and estimate the HR either in time domain or frequency domain.

The heartbeats are further affected by foot strikes during running (Figure 3c) which exhibit far stronger energy than any of the other activities, with high energy at 2.6Hz (Figure 3g) again corresponding to the cadence. The speech sound in Figure 3d also shows strong noise amplitudes due to the proximity of the ear and human vocal system, making the heartbeat-induced sound indiscernible. As in Figure 3f, the frequency components span over 1Hz to 4Hz and mask the HS, due to the jaw movement during speaking which creates vibrations and obscures the heart signals [19].

III. 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 HS 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 [20], [21]. 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 [22], [23], 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.

A. 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 V-B). 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 HS under motion. The discrete wavelet transform (DWT) is therefore used to remove artefacts from the audio to isolate HS. 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 HS 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 HS. 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 heart rate. Thus, we acknowledge that a more sophisticated approach to the problem, specifically to addressing signal denoising, is required.

B. Overview of the Deep Learning-Based Pipeline

An overview of hEARt, our designed motion-resilient HR monitoring system, is given in Figure 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 HS. For MA elimination,

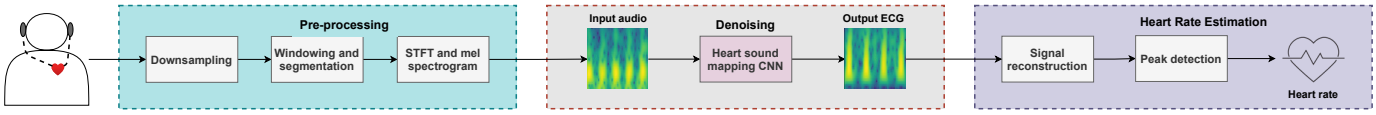


Fig. 4: hEARt system flowchart.

we proposed a CNN-based network to map spectrograms of the noisy HS signal to spectrograms of its 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 HS) is commonly represented in image form as spectrograms [24]–[26]. Initially developed for biomedical image segmentation, U-Net shows great potential in image denoising and super resolution [24], [27]. 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 [28], while complex network structures [29], [30] could easily suffer from overfitting. Finally, HR is estimated using peak detection on the clean signals.

C. Pre-processing

The HS 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 2s windows, each with a 1.5s overlap with the previous window. 2s 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 II-B, MAs and other interfering signals lie overlapping frequency ranges with HS, therefore requiring additional processing.

We process the GT similarly. The ECG, sampled at 130 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 2). Since we are only interested in capturing the beats and the inter-beat timing (for measuring HR, and in future, HRV), only the QRS complex is of interest.

D. Motion Artefact Elimination

1) Spectrogram Generation: 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. 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 HS frequencies reside. The resulting log-mel spectrogram is a 64x64 matrix for each window. Since audio is captured in both ears, a spectrogram is computed for each channel and stacked together to form one 64x64x2 input. The output is a single channel ECG spectrogram. The spectrograms are normalised between 0 to 1, to aid network training. Normalisation is carried out by dividing by a constant value, to maintain the difference in the signal amplitude for different activities.

2) Network Structure: Figure 5 provides the architecture of the denoising U-Net. In the encoder (or contraction path), the model consists of repeated 3x3 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 1x1 convolution is used to map the feature maps into a single 64x64 output image.

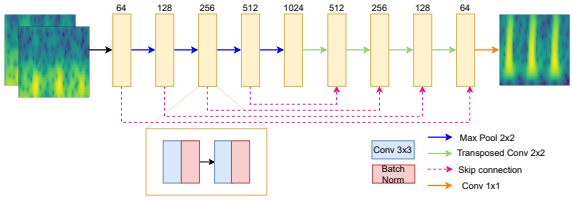


Fig. 5: U-Net autoencoder architecture.

3) Transfer Learning: On account of the small dataset, transfer learning is used to improve the results of the HS denoising. To achieve this, the model is pre-trained using the PASCAL HS dataset [31], where log-mel spectrograms of HS 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 HS. 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 HS using PASCAL, as well as avoiding overfitting on a small dataset.

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 64. When choosing training parameters, our objective was to strike a good balance between performance and computational complexity.

The system uses mean square error (MSE) or L2 loss (Equation (1)). This loss minimises the distance between the GT ECG spectrogram (y_{ij}) and the noisy audio spectrogram ($\hat{y}_{ij}$), where i and j represent the time and frequency index respectively, and T and F represent the total number of bins over the time and frequency dimensions respectively.

$$MSE = \frac{1}{TF}\sum(y_{ij} - \hat{y}_{ij})^2 \quad (1)$$

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.

E. Heart Rate Estimation

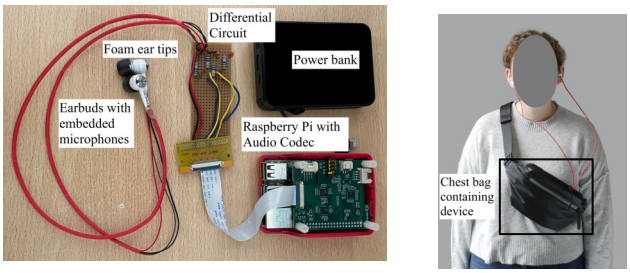
HR estimation is performed in an 10s long window, where each window has a 5s overlap with the previous window [33]. Each window undergoes the Hilbert transform to compute the envelope of the signal. Thereafter, a Gaussian moving average filter smooths out small ripples and peaks in the signal. Peak detection is calculated 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.

IV. 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.

A. Prototyping

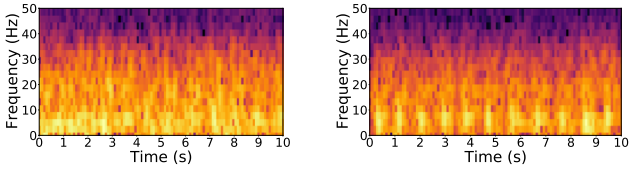
Although in-ear microphones have been integrated into existing commercial earbuds (e.g., AirPods Pro), no API is available to access the microphone output. To gather data and understand the potential of our approach, we developed our earbud prototype by customizing existing earbuds (Figure 6). Specifically, we embedded two analogue omnidirectional MEMS microphones (Knowles SPU1410LR5H-QB [15]) into a pair of wired earbuds, as shown in Figure 6a. The microphones were selected due to their flat frequency response from 10 Hz to 10 kHz which encompasses the frequency range of HS, speech and MAs. The microphones were connected to a differential circuit for common mode rejection of power line noise and other noise sources and sampled by an audio codec (ReSpeaker Voice Accessory Hat [34]) onto a Raspberry Pi 4B. For portability, the circuitry and Raspberry Pi were placed in a chest bag which was worn by participants during the experiments (Figure 6b). This ensured that the device did not interfere with the participant's natural movement during the tasks.



(a) (b)

Fig. 6: (a) Prototype and (b) participant wearing the device.

Although the occlusion effect implies the possibility of detecting bone-conducted sounds from the ear canal, measuring HS with an in-ear microphone is extremely challenging. Unlike walking, which generates strong vibrations, heart beat movement is subtle, resulting in very weak HS. As shown in Figure 7a, when using the earbud equipped with a silicon ear tip, it is difficult to identify heart beats from the signal. We overcome this challenge by replacing the silicon ear tip with a foam ear tip, which (1) largely suppresses/absorbs external sounds, resulting in a lower noise floor; (2) ensures a better sealing of the ear canal, thereby winning more amplification gain from the occlusion effect (shown in Figure 7b). With this upgrade, our prototype is able to measure HS with good SNR.



(a) (b)

Fig. 7: Comparison of signals collected when occluding the ear-canal with (a) a silicon ear tip and (b) a foam ear tip.

B. Data Collection

We used an ECG chest strap (Polar H10 [35]) to measure the GT heart signal. We extracted the raw ECG from the Polar H10 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 130 Hz. Due to the difference in the sampling rates, there is a maximum of a 150ms delay between the audio and the ECG signal. However, since HR estimation is performed in 10s windows, this delay is negligible. We synchronized the data by aligning the timestamps of the ECG signal with the timestamps of the audio file.

We invited 20 participants (13 males and 7 females) for data collection[1]. In addition to the stationary case, we considered three activities that are regarded as active or, that, because of their nature, interfere with the in-ear microphone: walking, running, and speaking. These activities were also selected as they match the conditions in [7] used to study in-ear PPG. While wearing our earbud prototype and the chest strap, the participants kept stationary for 30 seconds to obtain a reference HR. Then, they performed each of the tasks continuously for 2 minutes [36]. The 30 second stationary reference is only used for the baseline signal processing approach, and is not used in the hEARt system. When performing the walking and running activities, participants were allowed to pick a comfortable pace, and were instructed to move freely within a 5x4 m area. For the speaking activity, they were given a passage to read out loud. We processed a total of 160 minutes of in-ear audio corresponding to four tasks across all participants. Data collection was done in the atrium of a busy building, and as such data was collected in the presence of uncontrolled ambient noise, including human speech, the opening and closing of doors and low frequency power grid hum and air conditioning. The data collected while running for participants 2 and 14 was excluded on account of poor data quality. This occurred because one of the earbuds fell out during the intensive running activity. As such, there was no seal between the ear and the earbud meaning that the occlusion effect could not be leveraged.

The distribution of GT HR varies per activity. While stationary, the mean HR is 70±12 BPM, with a minimum and maximum of 45 and 114 BPM. While walking, the HR ranges from 51 to 129 BPM with a mean HR of 86±14 BPM. Running has the highest average HR (109±23 BPM) with the largest range of HR (50 to 187 BPM). The HR while speaking is similar to that while stationary with a mean of 76±12 BPM and a range of 51 to 124 BPM.

V. Performance Evaluation

A. 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 ($i, i \in [1, N]$).

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

(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 10s window.

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

B. Baseline Comparison

Table I shows the performance comparison between the proposed DL-based hEARt system and two signal processing approaches - (1) the proposed signal processing (SP) 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. Our proposed SP approach outperforms the baseline significantly for stationary and running, and marginally for walking and talking. This demonstrates that the baseline algorithm designed for stationary is unable to generalize to motion conditions, and an additional denoising module is required. Comparing the SP with hEARt, we observe that hEARt outperforms SP for each of the activities, showing that the DL based technique is better at generalizing to the differences in the data than the SP approach. While performance in the stationary case is comparable, with more intense motion interfering with the HS, SP fails to capture the HR from the signal and the performance severely deteriorates. hEARt outperforms SP significantly with a relative improvement of 51%, 54% and 48% for walking, running and talking respectively, suggesting the effectiveness of hEARt in HR estimation. Additionally, errors for the stationary, walking and running conditions are less than 10%, meaning that the system is accurate by ANSI standards for these activities [14].

The results for speaking are noticeably the worst of the four activities studied. This is consistent with Figure 3h, where it is clear that speaking brings more severe noises than the other activities. Perhaps against intuition, this 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 heart beats. They are also non-periodic and random in nature and are 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.

Table I 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.

C. hEARt Overall Performance

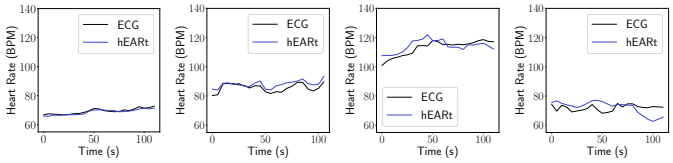
Figure 8 shows the qualitative assessment of hEARt in tracking HR over time. We compared the GT HR collected via an ECG chest-strap with the one extracted from the in-ear audio for one participant over the four different activities. It can be observed that the proposed approach is able to accurately

TABLE I: 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	**4.32 ± 3.99**	4.93 ± 8.33	9.88 ± 6.93	—
Walking	**9.53± 8.28**	19.41 ± 16.03	20.90 ± 11.22	27.14
Running	**9.80 ± 7.93**	21.43 ± 15.30	34.28 ± 8.73	29.84
Speaking	**12.06 ± 8.88**	23.37 ± 9.39	24.23 ± 7.98	12.52

and continuously track the user's HR during the four activities (stationary, walking, running, and speaking), suggesting the potential of in-ear audio for HR estimation under MA. For speaking, the larger error is due to jaw movements. However, the overall trend of estimated HR still aligns with GT.

Overall, the system achieves a MAE of 3.02 ± 2.97 BPM, 8.12 ± 6.74 BPM, 11.23 ± 9.20 BPM and 9.39 ± 6.97 BPM for stationary, walking, running and speaking respectively. As noted, we achieve the lowest performance during speaking in terms of MAPE as shown in Table 2. Concretely, given an average heart rate of 76 BPM (the mean HR while talking as per Section IV-B), a MAPE of 12.06% means our system misses (or adds) about 0.15 heart beats every second, or, 1 heart-beat every 7 seconds. Similarly, the performance achieved for running is even more convincing: at an average heart rate of 109 BPM, we miscompute 0.18 heart beats per second, amounting to around 1 heart-beat every 5 seconds.



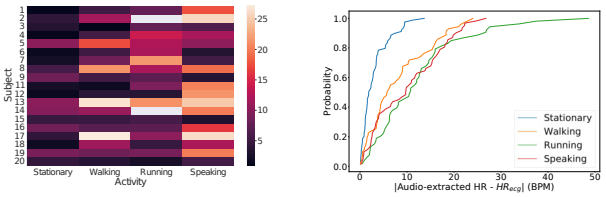
(a) Stationary (b) Walking (c) Running (d) Speaking

Fig. 8: Qualitative longitudinal performance of heart rate extraction under different activities.

D. Individual HR Estimation

Next, we evaluate our approach under different activities for all subjects. First, we provide some insights on the population statistics. Figure 9a reports a heatmap of the MAPE of the audio-extracted HR for every user across the activities. Lighter colors correspond to greater MAPE values. Running for user 2 and 14 was removed due to a poor seal, and is represented as a white 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. **(iii)** One user experiences overall poor performance (user 13). This is due to a poorly fitting earbud, and poor quality GT data. **(iv)** Certain users experience poor performance in a specific activity (e.g. user 17 for walking). This is again 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 a higher quality earbud fit. 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.

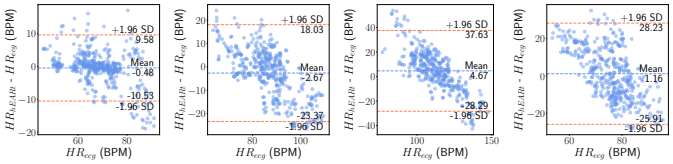
To further understand the extent to which the various activities impact hEARt, for each of them we report the empirical cumulative distribution function (ECDF) of the error (Figure 9b). Looking at the ECDFs we can confirm what was observed in the heatmap. Specifically, our approach achieves an error of less than 12 BPM for over 60% of users for all activities. As seen in the heatmap, most of the error observed comes from a few specific users rather than from the population in general. This performance on our academic prototype confirms that in-ear audio sensing of HR offers a promising alternative for continuous HR sensing in presence of motion.



(a) (b)

Fig. 9: (a) MAPE heatmap per subject and (b) empirical CDF.

E. Bland-Altman Plots



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

Fig. 10: Modified Bland-Altman plot of heart rate extraction.

To further analyze the results, we leverage modified BA plots. We report the BA plots (i.e. the agreement between the HR calculated with hEARt and that obtained from the GT chest strap) for each condition in Figure 10. Specifically, Figure 10a reports the agreement while stationary. It is clear that the bias between the two measurements is minimal, with very low mean (only -0.48 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 10b and Figure 10c respectively), wider limits of agreement are present, representing a greater standard deviation in the HR estimation. Interestingly, while overall the mean errors remain low (-2.67 BPM for walking and -2.41 BPM for running), our approach exhibits a larger error for estimation as HR increases. We observed this phenomenon both for walking (Figure 10b) and running (Figure 10c) motions. Notably, especially in the running case, this is observed when the frequency of the running overlaps with the HR values. The spurious MA-induced spikes trigger a harsher response

by hEARt that tries to remove the noisy peaks, thus leading to an underestimation of HR above 120 BPM. Additionally, another factor to explain the higher errors biased towards higher heart rates could be traced back to the imbalance of our dataset, where lower HR values are predominant. Finally, in Figure 10d, the mean error is again low but with fairly wide limits of agreement. This wide standard deviation again points towards the complexity of the speaking activity, meaning that extrapolating useful heart signals to compute HR from in-ear audio signals is a very challenging task, never tackled before. Nonetheless, our approach still performs well.

F. Long-term tracking performance

The results of the previous sections were obtained from experiments run under controlled conditions. To assess the real world effectiveness of the designed system, we collected an hour of data from one subject under conditions of daily life. During this time, the subject was instructed to undergo their activity as normal. This activity included working in an office, walking around, speaking (while working) and taking a short jog. The results of HR prediction for this study are given in Figure 11. 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 closely match. However, as was seen in the BA plots in Section V-E, the system underestimates the higher heart rates. This is likely due to the distributions of heart rates in the dataset where the average HR is 85 BPM.

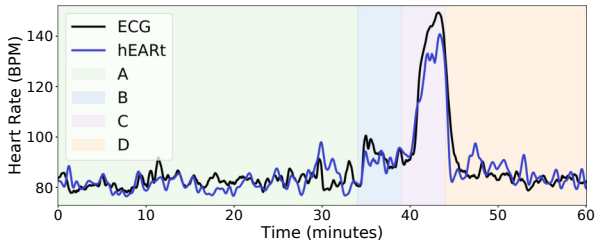


Fig. 11: Longitudinal heart rate tracking. Coloured boxes indicate the different activities. A: Working while sitting. B: Walking. C: Running. D: Working while standing.

The MAE of this longitudinal study is 4.96 BPM, which is a MAPE of 5.34%. To further break this down, the MAPE of activities A, B, C and D are 4.41%, 5.59%, 9.51%, and 5.94% respectively. If we compare these results to those in Table I, we can see that all 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. It also shows that the model is able to make accurate predictions even under conditions of mixtures of activities. Thus this study acts as a proof of concept of the in-the-wild feasibility of the hEARt system.

G. Power and Latency Measurements

To provide a full system analysis, we assess the power consumption and latency of the system implemented on a Raspberry Pi4. 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 II provides a breakdown of the operation times for the various system components. Signal denoising was performed on a 2s window and HR extraction on a 10s window, as detailed in Section III. Processing a 10s window takes the system 54.35ms, implying that a new HR can be predicted by the system every 5s (due to the 5s second overlap). This latency is an adjustable parameter of the system based on the overlap ratio. The system power consumption is given in Table III. Overall, the system (including microphone sampling, denoising and HR prediction) consumes 701mW. The microphone sampling runs continuously, but the hEARt system is only active for 65.04ms for each estimate, and an estimate is made every 5s. Thus, the average energy consumed per second is $(2871-2775)mW \times 1s + (3547-2871)mW \times 65.04ms/5 = 104.79mJ$. In context, if run on a wireless earbud such as the Apple Airpod Pro (with a battery capacity of 43mAh[2]), hEARt could operate continuously for a time $T = \frac{43mAh \times 5V}{104.79mJ} = 2.05hr$. While this may seem like a short operating time, this system has been implemented on a power hungry Raspberry Pi without optimizing for energy consumption. 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. However, ultimately this gives an indication of how hEARt could feasibly be implemented on a commercial earbud.

TABLE II: Latency.

Operation (window)	Latency (ms)
Preprocessing (2s)	1.66
Denoising (2s)	7.66
Reconstruction (10s)	17.96
HR extraction (10s)	0.48
Total (10s)	65.04

TABLE III: Power consumption.

Operation	Power (W)
RasPi (Baseline)	2.775
RasPi+Mic	2.871
Full system	3.547

VI. 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 we use [15] has a current draw of $0.12mA$, more than 10 times less than that of a state-of-the-art wearable dedicated PPG module, the MAXM86161, which draws $1.62mA$ [16].

In the remainder of this section we reason over some shortcomings of our work, and potential solutions. First, we

[2]https://www.ifixit.com/Teardown/AirPods+Pro+Teardown/127551

are aware of the limitations that come with a simple, cheap prototype like ours. For instance, some of the collected data was corrupted as the subjects were unable to wear the earbuds properly, even though they were asked to fit them tightly. 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]. The data corruption while running was also worsened by the wires on the earbuds which move during vigorous activity thus dislodging 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 prototype.

We note that the MAPE in HR estimation while speaking requires improvement to meet ANSI standards for HR monitors. Since the system aims to determine HR under active conditions (e.g., running), we expect the amount of speaking to be less than in non-active scenarios, limiting the impact of those errors especially over prolonged periods. The errors occur since speaking introduces non-stationary noise that is different than walking and running. Other techniques to remove non-stationary noise can be considered, or the quantity of speaking samples could be increased during training so that the model can learn characteristics of these signals better.

Given that earbuds are mainly used for audio delivery, one concern is whether music playback will affect performance. As studied by Ma et al. [10], in music, the average energy ratio below 50 Hz is only 1.5%. This means that music has a negligible impact on our system as it operates on signals below 50 Hz. To confirm this, we first superimpose music on the collected in-ear signals and filter it with a lowpass filter (pass band $<$50Hz). We then compute the Pearson correlation between the filtered signal and original signal, yielding a coefficient of 0.982, further proving our system's robustness.

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 HS. Additionally, collecting data from more subjects encompassing a wider range of HR will improve the ability of the model to further generalize to higher HR.

VII. Related Work

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 [39], dietary monitoring and swallow detection [40], blood pressure monitoring [41], step counting [10], heart and respiratory rate tracking [8], user identification and gesture recognition [42], etc. A paradigm named HeadFi was proposed to turn the drivers inside existing headphones into a sensor, with its potential validated in four applications [42]. 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, or speaking.

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. 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%, 29.84% and 12.52% error of PPG sensors in earables for walking, running and speaking respectively, quantitatively demonstrating the challenges of PPG in HR estimation under motion. Acoustic sensors have also been studied for HR measurements. Chen et al. [43] estimated HR from a small acoustic sensor placed at the neck. [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. [44] 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 HS is not investigated. These findings imply the challenges in HR measurement from earables under motion. We have thus presented an approach that aims to tackle these and offer a solution to measuring HR in realistic settings.

VIII. Conclusion

We proposed an approach for accurate HR estimation using audio signals collected in the ear canal, under motion artefacts caused by daily activities (e.g., walking, running, and talking). 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 3.02 $\pm$ 2.97 BPM, 8.12 $\pm$ 6.74 BPM, 11.23 $\pm$ 9.20 BPM and 9.39 $\pm$ 6.97 BPM for stationary, walking, running and speaking, respectively, opening the door to new non-invasive and affordable HR monitoring with usable performance for daily activities. We also discussed some potential strategies to further improve the performance in the future.

REFERENCES

[1] 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.

[2] 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, 12 2020.

[3] 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.

[4] 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*, vol. 8, pp. 184 774–184 784, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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, 1 2014.

[10] 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, p. 175–187.

[11] M. Murray, G. Spurr, S. Sepic, G. Gardner, and L. Mollinger, "Treadmill vs. floor walking: kinematics, electromyogram, and heart rate," *Journal of applied physiology*, vol. 59, no. 1, pp. 87–91, 1985.

[12] 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.

[13] 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.

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

[15] *Zero-Height SiSonic Microphone With Extended Low Frequency Performance*, Knowles Electronics, 3 2013, rev. D.

[16] *Single-Supply Integrated Optical Module for HR and SpO2 Measurement*, Maxim Integrated, 3 2019, rev. 0.

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

[18] 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.

[19] A. Bedri, D. Byrd, P. Presti, H. Sahni, Z. Gue, and T. Starner, "Stick it in your ear: Building an in-ear jaw movement sensor," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 2015, pp. 1333–1338.

[20] M. Ali and 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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. [Online]. Available: https://arxiv.org/abs/1802.06955

[29] J. Yoon, D. Jarrett, and M. van der Schaar, *Time-Series Generative Adversarial Networks*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[30] Q. Man, Y.-I. Cho, S.-G. Jang, and H.-J. Lee, "Transformer-based gan for new hairstyle generative networks," *Electronics*, vol. 11, no. 13, 2022.

[31] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, "The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results," http://www.peterjbentley.com/heartchallenge/index.html.

[32] 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.

[33] 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.

[34] "ReSpeaker 4-Mic Linear Array Kit for Raspberry Pi - Seeed Wiki," https://wiki.seeedstudio.com/ReSpeaker_4-Mic_Linear_Array_Kit_for_Raspberry_Pi/, accessed Oct 10, 2022.

[35] "Polar H10 — Polar Global," https://www.polar.com/en/sensors/h10-heart-rate-sensor, accessed Oct 10, 2022.

[36] M. Almeida, A. Bottino, P. Ramos, and C. G. Araujo, "Measuring Heart Rate During Exercise: From Artery Palpation to Monitors and Apps," *International Journal of Cardiovascular Sciences*, vol. 32, pp. 396–407, Aug. 2019.

[37] S. Ismail, U. Akram, and I. Siddiqi, "Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review," p. 5, 12 2021.

[38] D. Giavarina, "Understanding Bland Altman analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[39] 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, 5 2014.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.