EarMeter: Continuous Respiration Volume Monitoring with Earables

YANG LIU, University of Cambridge, United Kingdom
QIANG YANG, University of Cambridge, United Kingdom
KAYLA-JADE BUTKOW, University of Cambridge, United Kingdom
JAKE STUCHBURY-WASS, University of Cambridge, United Kingdom
DONG MA, Singapore Management University, Singapore
CECILIA MASCOLO, University of Cambridge, United Kingdom

Respiration volume, i.e., the amount of air inhaled/exhaled during breathing, is a critical measure for health and fitness in daily life, such as helping optimize sports performance, tracking wellness, and early anomaly detection. Current continuous respiration volume monitoring solutions either require specialized and cumbersome instrumentation setup (e.g., RF transceivers), or rely on customized and non-portable wearables (e.g., masks and chest straps), limiting their usage scenarios. In this paper, we introduce EarMeter, the first continuous respiration volume monitoring system that utilizes in-ear microphones on earbuds to seamlessly track respiration volume across varying breathing intensities, making the measurement more accessible in diverse scenarios. The underlying idea is that breathing sounds, which correlate with breathing volume, can propagate through the body to the ear canals, where they are captured by in-ear microphones. To achieve this, we propose a deep-learning approach to address four unique challenges: limited labeled data, faint breathing sounds, interference from footsteps, and generalization to unseen users. Our approach features fine-tuning an audio encoder pretrained on a broad range of audio datasets, knowledge transfer from high-quality nose audio, performance boosting with breathing-heartbeat coupling, and alignment of both earphone channels with normalization. Extensive experiments under the Leave-One-Subject-Out (LOSO) setting across varying breathing intensities demonstrate the effectiveness of EarMeter, with an average Mean Absolute Percentage Error (MAPE) of 18.19%, meeting the clinically required standard of 20%.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**.

Additional Key Words and Phrases: Earable Sensing, Respiration Monitoring, Respiration Volume, In-ear Microphone

ACM Reference Format:
Yang Liu, Qiang Yang, Kayla-Jade Butkow, Jake Stuchbury-Wass, Dong Ma, and Cecilia Mascolo. 2025. EarMeter: Continuous Respiration Volume Monitoring with Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 198 (December 2025), 28 pages. https://doi.org/10.1145/3770655

1 Introduction

The assessment of respiratory function plays a crucial role in the understanding of respiratory conditions and the fitness level of the human body [27, 56, 73]. Recent studies have primarily focused on respiratory rate monitoring [25, 44, 51, 67, 76], which is useful but not fully indicative of lung conditions, as it does not provide information about the air exchanged in the lungs. *Respiration volume*, the amount of air inhaled and exhaled during respiration, is an important biomarker describing an individual's ventilatory status, providing valuable

Authors' Contact Information: Yang Liu, yl868@cam.ac.uk, University of Cambridge, United Kingdom; Qiang Yang, qiang.yang@cl.cam.ac.uk, University of Cambridge, United Kingdom; Kayla-Jade Butkow, kjb85@cam.ac.uk, University of Cambridge, United Kingdom; Jake Stuchbury-Wass, js2372@cam.ac.uk, University of Cambridge, United Kingdom; Dong Ma, dongma@smu.edu.sg, Singapore Management University, Singapore; Cecilia Mascolo, cm542@cam.ac.uk, University of Cambridge, United Kingdom.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2025 Copyright held by the owner/author(s).
ACM 2474-9567/2025/12-ART198
https://doi.org/10.1145/3770655

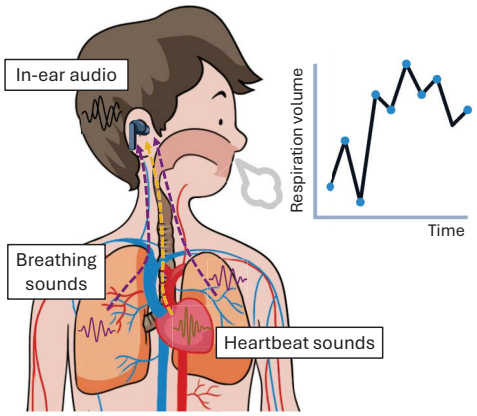


Fig. 1. EarMeter utilizes in-ear microphones on earphones to capture both breathing sounds and heartbeat sounds, enabling reliable continuous breathing volume estimation across varying breathing intensities.

insights into a person's overall health, well-being, and everyday lifestyle [48]. In fitness and sports science, understanding respiration volume helps optimize performance and monitor the recovery process [52]. Regularly measuring respiration volume provides a picture of general health as you age [7], and significantly lower values can act as a warning for underlying problems that warrant further attention from healthcare providers [9]. Breathing volume also reflects our mental states in everyday life, including fatigue and even stress or anxiety levels [71]. As such, **a portable solution for continuously monitoring breathing volume in daily life is important to improve public health and wellness**.

Recent works [67, 74] focus on spirometry tests that estimate lung capacity through one-time respiration volume estimation, where users are required to take a deep breath and then exhale as forcefully and completely as possible. On the contrary, continuous respiration volume monitoring—which captures natural, ongoing breathing patterns to continuously track lung function and detect subtle health changes—is highly desirable. However, it faces several practical barriers. Current gold-standard methods for continuous respiration volume monitoring, such as plethysmography [20], are primarily employed in clinical settings. These methods require individuals to visit a hospital or clinic, where measurements are conducted under the supervision of medical professionals using bulky equipment costing thousands to tens of thousands of USD. In recent years, wearable devices for continuous respiration volume monitoring have started to be investigated, including sensors embedded in fitting garments or elastic bands [8, 39, 59, 64] or facial masks [4, 83]. However, these methods are cumbersome, uncomfortable, and socially unacceptable for everyday use.

Earphones are increasingly popular and are often worn during exercise, work, and leisure activities [68, 69, 72, 77]. Recently, their on-board sensors have enabled a preliminary study [61] on continuous breathing volume estimation. In [61], photoplethysmography (PPG) in earphones was used to estimate respiration volume, but the method was limited to a constrained scenario where users remained still and their respiration rate was within a fixed range. In addition, this approach resulted in a high MAPE of nearly 40%. To the best of our knowledge, **there currently exists no low-cost, non-intrusive, portable, and accurate solution for continuous breathing volume estimation in daily life**.

In this paper, we introduce EarMeter, a novel respiration volume monitoring system that enables continuous monitoring using in-ear microphones on earphones, seamlessly across varying breathing intensities and daily activities. In-ear microphones, commonly found in earphones with active noise cancellation (ANC), make

continuous respiration volume tracking more accessible, scalable, and practically possible. As depicted in Figure 1, breathing sounds are generated by airflow turbulence in the respiratory system during inhalation and exhalation, which correlates with the volume of respiration. These sounds then travel through bones and tissues to the ear canal, where they are captured by the in-ear microphones and used by EarMeter for respiration volume monitoring. In this work, we focus on estimating respiration volume in healthy individuals under daily living conditions, providing a foundation for future studies in broader populations. Developing EarMeter involves several challenges:

(1) The latent mapping between in-ear breathing signals and respiration volume is complex and inherently non-linear. However, the limited paired in-ear audio and respiration volume make it challenging to model this relationship. To address this issue, we tune an audio encoder pretrained on a broad range of sounds with a small amount of our task-specific data, allowing it to learn robust audio representations.
(2) The breathing sounds captured in the ear canal become extremely weak when the user is at rest (with natural, low-intensity breathing rhythms). To address this issue, we observed that the energy of heartbeat sounds captured by in-ear microphones remains consistent in such cases. Inspired by the coupling relationship between the cardiovascular and respiratory systems (as detailed in Section 4.4) [80], we explicitly extract both heart and breathing features from the in-ear audio and exploit this phenomenon to enhance respiration volume estimation.
(3) In-ear breathing sounds are easily overpowered by other bone-conducted noises, such as footsteps when the user is moving. Naively tuning a pretrained encoder using data containing such interference would likely yield a model capturing the characteristics of high-energy footstep sounds rather than breathing sounds, severely degrading performance. On the other hand, excluding data with interference from the training process would also yield a model that cannot reliably estimate breathing volume during movement. To tackle this challenge, we propose a novel knowledge transfer framework that utilizes high-quality nasal audio to guide the model in learning effective breathing features from the in-ear audio, even with interference.
(4) Given the *inter-individual* variability in physiological characteristics, body types, and signal patterns, generalizing to unseen users often requires fine-tuning with user-specific data to achieve satisfactory accuracy. However, this process typically depends on access to clinical facilities and specialized ground-truth devices, which can be bulky, expensive and may limit practical usage. Moreover, *intra-individual* variability also exists due varying placements of earbud on each ear. To address these concerns, we 1) introduce a feature alignment framework we call "earphone channel alignment" that pushes features extracted from both earphone channels into the same learned embedding to reduce intra-user variability resulting from earbud placements, and 2) incorporate a normalization component to mitigate inter-user variability. Our experiments show that EarMeter achieves an average MAPE below 20%, aligning with clinical standards. Incorporating these techniques nearly doubles the number of new users under this threshold compared to without them, providing promising evidence that a generalizable model that requires little to no fine-tuning with medical-grade devices is feasible.

To evaluate the performance of EarMeter, we developed an earable prototype and conducted experiments with 22 healthy participants across varying breathing intensities under both stationary and moving scenarios. EarMeter achieves an average Mean Absolute Error (MAE) of 0.20 liters (L), MAPE of 18.19%, and Pearson Correlation of 0.89 in a leave-one-subject-out (LOSO) validation, which meets the clinical standard of a MAPE of less than 20% [65]. We also evaluate EarMeter's performance on the other two biomarkers of the respiration volume at different scales, showing the importance of the approach and its wide applicability.

In summary, this paper makes the following contributions:

- We introduce EarMeter, the first *earable-based* system for *continuous* respiration volume monitoring that utilizes in-ear audio across varying breathing intensities. Unlike existing works that are constrained to limited

settings, EarMeter enables continuous respiration volume monitoring across a wide range of scenarios in daily life that was previously not possible.
- We propose a novel deep learning approach to tackle four key challenges in EarMeter: limited labeled data, faint breathing sounds, interference from footstep noise, and generalization to unseen users. Our framework introduces (i) a knowledge transfer strategy that leverages representations learned from a small amount of high-quality audio collected under the nose, (ii) an exploitation of the coupling between cardiovascular and respiratory systems to improve respiration volume estimation, and (iii) an earphone channel alignment with normalization method to mitigate inter- and intra-user variability.
- We develop an EarMeter prototype and conduct extensive benchmarks involving 22 subjects across varying breathing intensities. The results demonstrate that EarMeter can achieve a MAPE of 18.19%, which surpasses the clinical standard of a MAPE of 20% or less, even under motion and in daily living conditions. Deploying EarMeter on a mobile phone also demonstrates real-time performance with low energy consumption.

2 Related Works

2.1 One-time Spirometry Test

Existing studies [67, 74] have explored respiration volume estimation during spirometry tests for monitoring human lung function. These tests measure various volumes, including forced expiratory volume in one second (FEV1), forced vital capacity (FVC), and forced inspiratory vital capacity (FIVC). SpiroSonic [67] utilizes commodity smartphones to track chest wall motion through acoustic sensing during spirometry tests. Similarly, Earspiro [74] uses earphone microphones to estimate the flow-volume (FV) curve by analyzing airflow sounds recorded during these tests to estimate FEV1, FVC, and FIVC. A recent study [15] explored a similar concept but allowed participants to perform submaximal, rather than maximal, breathing maneuvers. However, spirometry tests require users to take a deep breath or exhale as forcefully and completely as possible, which makes them unsuitable for continuous, natural respiration volume monitoring [5, 67, 74]. Our approach diverges by focusing on continuous breathing volume estimation that captures the user's normal, everyday breathing patterns, as opposed to the one-time or (sub)maximal breathing volumes measured during traditional spirometry tests.

2.2 Continuous Respiration Volume Monitoring

2.2.1 Contact-free approaches: Contact-free systems utilizing RF signals [50, 51, 82] or cameras [51, 75] have been investigated for continuous respiration volume monitoring. DeepBreath [75] focuses on lung volume estimation during belly breathing for breathing exercise assessment. In their approach, the user is required to ensure that the hand on the belly exhibits regular upward and downward movements synchronized with the breathing cycle, while the hand on the chest remains relatively still. A forward-facing camera is used to capture these motions, allowing DeepBreath to estimate lung volume during a one-time exercise session with participant cooperation. In contrast, our system, EarMeter, is designed to continuously estimate breathing volume in natural settings without requiring user effort. WiKiSpiro [50] and WiSpiro [51] present a hybrid radio-camera system and a directional radio system, respectively, for estimating respiration volume during sleep. MoRe-Fi [82] introduces a motion-robust respiration monitoring system for waveform recovery. They use the amplitude of the recovered waveform to represent respiration volume, based on their proportional relationship. However, this method cannot provide absolute volume measurements, i.e., the exact number in liters. Additionally, these systems require extensive room instrumentation and setup, and their operational range is limited. In contrast, EarMeter offers a portable solution for absolute respiration volume monitoring using earphones, making it more accessible for everyday use and suitable for use beyond just resting scenarios.

2.2.2 Wearable-based approaches: Respiration volume has been explored using wearable devices as well; however, existing solutions typically rely on obtrusive sensors on the body, such as chest/upper-arm straps [8, 22, 39, 40,

64, 66], fitted garments [10, 14, 31], masks [4, 83], or direct attachment of multiple sensors to the skin [18, 19, 47]. These approaches can be uncomfortable and cumbersome, which significantly limits their acceptance for everyday use. A preliminary study presented at a workshop, OptiBreathe [61], explores estimating respiratory volume using PPG sensors in earphones. However, this approach was tested under constrained conditions—limited to a breathing rate of 10-20 bpm and stationary users, and their simple algorithm results in a substantial error (MAPE) of nearly 40%. Furthermore, while microphones are commonly integrated into earphones [12, 35, 36, 78], PPG sensors would need to be specifically added for monitoring purposes and are not yet a standard feature in most commercial earables. In contrast, EarMeter is the first continuous respiration volume monitoring system that utilizes earphone microphones to deliver promising performance across a range of breathing intensities.

3 Background and Challenges

3.1 Respiration Volume

Respiration volume, or *the average amount of air inhaled or exhaled per breath during natural breathing, computed over a specific period*, is a key indicator for assessing several dimensions of lung function, such as:

- **Obstructive breathing patterns**: These involve difficulty in exhaling air due to increased airway resistance, leading to slower respiration volume. They are important in sports science, where respiratory efficiency can influence athletic performance [53] and may reflect exercise-induced bronchoconstriction (EIB) experienced by endurance athletes [28].
- **Restrictive breathing patterns**: These occur when lung expansion is limited, hindering airflow and reducing respiration volume. They are observed in everyday situations such as shallow breathing during stress or anxiety, and in sleep-related conditions that affect breathing rhythms [54, 71].
- **Effective ventilation:** Essential for eliminating carbon dioxide (CO_2) and absorbing oxygen (O_2), effective ventilation is crucial during physical activities where CO_2 production increases and more O_2 is needed. In the general population, respiration volume increases naturally to meet these demands, making it a valuable indicator for monitoring physical exertion, cardiorespiratory fitness, and recovery [11].
- **Gas exchange efficiency:** When ventilation is insufficient, it can lead to elevated CO_2 levels (hypercapnia) or low oxygen availability (hypoxia) [21]. Tracking respiration volume continuously can provide insights into everyday scenarios such as fatigue, the influence of certain medications, or the impact of lifestyle factors like obesity on breathing patterns [63, 70].

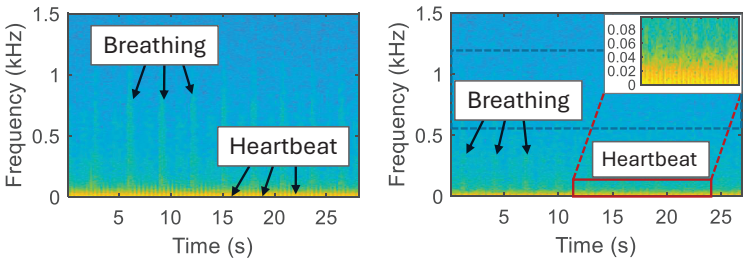
3.2 Breathing Sounds and Respiration Volume

Breathing sounds are produced by the airflow in the respiratory system during inhalation and exhalation. When air flows through the respiratory tract, it encounters resistance, leading to turbulence within the trachea and large airways (bronchi) [49]. This turbulence causes breathing sounds. The characteristics of these sounds can vary based on the respiratory cycle, the lung capacity, the velocity of the airflow, and the physical condition of the respiratory pathways [49]. The relationship between respiration volume V and the spectral power E of the breathing sounds $x(t)$ has been studied and is often modeled using a power-law equation [26, 60, 79]:

$$\begin{aligned} V &= C1 \times log(E) + C2 \\ E &= \int X(\omega) d\omega \end{aligned} \tag{1}$$

where $C1$ and $C2$ are the model coefficients related to individual physiological factors. $X(\omega)$ is the Fourier transform of $x(t)$ in the frequency domain.

As shown in Figure 1, the generated breathing sound propagates through the thoracic cavity, surrounding bones and tissues, and eventually reaches the ear canals, where it can be captured by in-ear microphones [45].



(a) High breathing intensity (b) Low breathing intensity

Fig. 2. An illustration of in-ear physiological sounds across different breathing intensities.

We can model the propagation of in-ear breathing sounds as follows:

$$y(t) = x(t) * h(t) + n(t) \tag{2}$$

where $y(t)$ is the double-channel breathing sounds captured in the ear canal at time t, $h(t)$ is the propagation channel and from the lung to the ear canal, and $n(t)$ represents noise. $*$ is the convolution operation. Figure 2a illustrates the spectrogram of in-ear physiological sounds, and we can clearly observe the breathing sound. Substituting Equation 1 into Equation 2, we can obtain the model from the in-ear breathing sounds to the respiration volume:

$$V = C1 \times \log\left(\int \frac{Y(\omega) - N(\omega)}{H(\omega)} d\omega\right) + C2 \tag{3}$$

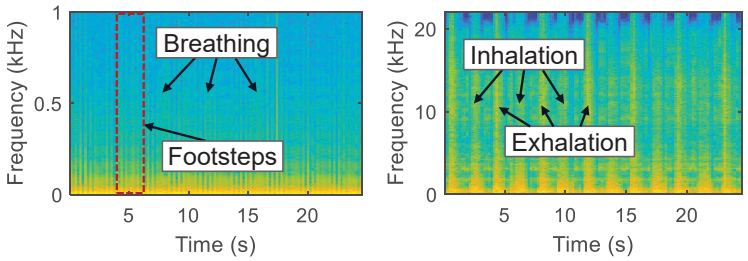
3.3 Challenges and Opportunities

Directly deriving respiration volume from in-ear sounds alone remains challenging due to the non-linear distortions caused by the complex interaction of physiological factors and sound propagation through the body that are difficult to model in Equation 3. To address this, *we leverage deep learning to approximate the intricate relationship between the spectral features of in-ear breathing sounds and respiration volume.* However, three major challenges remain:

1) Limited availability of labeled in-ear audio data. To the best of our knowledge, no existing datasets provide paired microphone and respiration volume data. Furthermore, collecting a sufficient number of labeled samples from a diverse participant pool is labor-intensive.

- *Fine-tuning an audio encoder.* Rather than training models entirely from scratch, we opted to *fine-tune an audio encoder pretrained on a broad range of audio datasets*, as discussed in **Section 4.2**. By leveraging an encoder that has been pretrained to interpret a wide variety of sounds, we can adapt our architecture to achieve robust performance with relatively small amounts of task-specific data, as demonstrated in Section 6.3. However, fine-tuning a pretrained model alone is still insufficient due to the following challenges.

2) In-ear breathing sounds are low in volume. In Figure 2a, we can clearly observe breathing sounds in the high-frequency band when a user engages in high-intensity breathing. However, when a user is at rest (with natural, low-intensity breathing rhythms), the breathing sounds become very weak. As shown in Figure 2b, after significant absorption and attenuation by soft tissues, the breathing sounds reaching ear canals become very faint, particularly in the high-frequency range (highlighted in the dashed box). This attenuation poses a challenge for accurate breathing volume estimation. To address this challenge, we leverage the following novel insight to boost performance:



(a) In-ear sounds. (b) Nasal breathing sounds.

Fig. 3. The audio signals captured in the ear canal and under the nose during running.

- *Boosting performance with heart sounds.* One pattern we observed is that the energy of heartbeat sounds captured by in-ear microphones are often more prevalent than breathing sounds when a user is at rest, and generally do not overlap with the breathing sounds, as shown in Figure 2. Past work established physiological couplings between the respiratory and cardiovascular systems [80], which introduces specific modulations in heartbeat signals that are closely tied to respiration. In this work, we exploit these couplings to enhance respiration volume estimation when reliable heartbeat sounds are present in the ear (**Section 4.4**).

3) In-ear breathing sounds can be easily overpowered. In real-world scenarios, users are often moving around (e.g., walking or running) and not stationary or standing still. Figure 3a shows the spectrogram of in-ear audio while a person is running, which clearly overwhelms most of the breathing signal. Directly tuning a pretrained encoder using data contaminated with such interference would likely yield a model that captures the characteristics of high-energy footstep sounds, rather than breathing sounds, leading to severely degraded performance. On the other hand, removing data with interference from the training process would also yield a model that cannot reliably estimate breathing volume during movement. To address this challenge, we leverage two novel insights to boost performance under noisy scenarios.

- *Learning respiration volume from the nose.* The best area on the body to learn breathing volume from sound is the area with the highest breathing signal-to-noise (SNR) ratio, since this location best captures the feature (breathing sounds) that we are using to translate to breathing volume. We found that breathing sounds are clearly audible with a microphone placed just below the nose. As shown in Figure 3b, these nasal breathing sounds are strong and unaffected by footstep noise, distinctly capturing both inhalation and exhalation patterns. Therefore, we decide to create a representation that can accurately estimate breathing volume from a microphone placed under the nose (**Section 4.3.1**).
- *Knowledge transfer from nose to ear.* Estimating breathing volume from a microphone under the nose requires creating a custom wearable that specifically places the microphone at that location. Such a device is not commonplace, nor is natural to wear. However, the sounds generated through the nose is directly correlated with the breathing sounds measured in-ear, with the major difference being the channel that the sound propagates through (e.g., nose: nasal cavity, ear: bone, skin, and tissue). To leverage the representations learned from the nose, we propose a novel knowledge transfer framework, where the in-ear respiration volume model, acting as the student network, leverages the nose model, acting as the teacher network, to better learn weights, representations, and predictors of breathing volume (**Section 4.3.2**).

4) Inter-individual and Intra-individual variability requires additional calibration. Generalizing respiratory volume estimation models to unseen users remains a significant challenge due to the wide inter-individual variability in physiological characteristics, such as lung capacity, chest wall compliance, and breathing habits. In addition, variations in ear canal shape, body composition, and sensor placement can further influence the

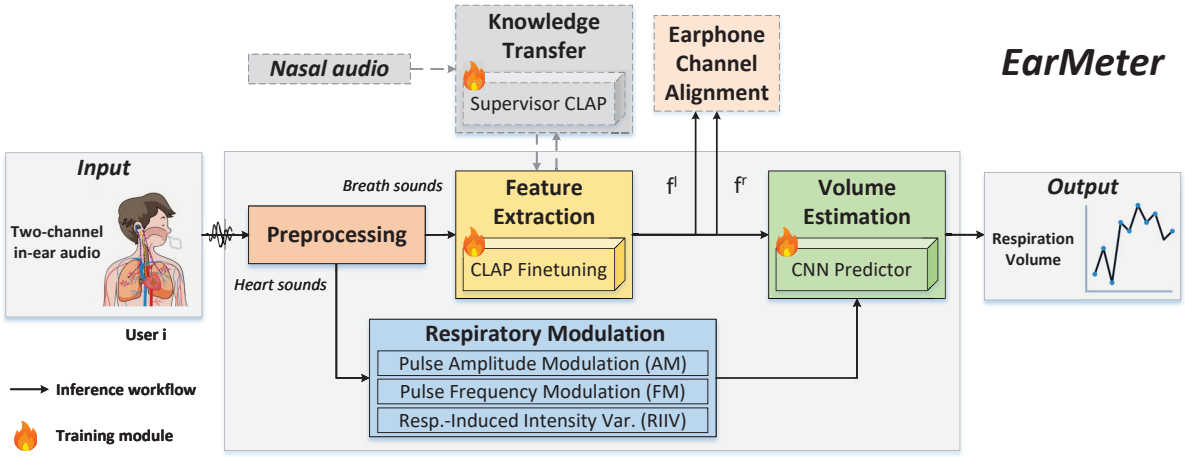


Fig. 4. Illustration of the EarMeter pipeline. f_l and f_r refer to the intermediate features extracted for the left and right in-ear microphone channel, respectively.

acoustic and inertial signals captured by earable devices. As a result, models trained on one group of users often require fine-tuning with user-specific data to maintain satisfactory accuracy on new individuals. However, this fine-tuning process typically requires access to clinical equipment (e.g., spirometers, metabolic carts, or high-end masks) for label data collection, which typically require clinical infrastructure and/or trained personnel, making them hard to access. In addition to variability between users, variations in the placement of the right and left earbuds also introduce differences in sound occlusion, signal characteristics, and channel intensities even for the same individual. For example, the degree of occlusion may vary depending on how far an earbud is inserted into the ear canal and its orientation. To address these limitations, we introduce two key components (**Section 4.5**).

- *Label Normalization.* First, we apply global min–max normalization to the respiration volume labels so that all subjects are mapped to a common scale. This reduces inter-subject variability caused by differences in lung capacity, stabilizes model training, and prevents the model from being biased toward subjects with naturally larger volumes.
- *Earphone Channel Alignment.* Second, to account for intra-user variations caused by differences in earbud placement between ears, we introduce an alignment scheme that encourages EarMeter to learn consistent embeddings from the right and left microphone channels for the same breathing event. This helps EarMeter filter out and compensate for deviations between in-ear microphones within the same user.

4 System Design

4.1 System in a Nutshell

Figure 4 shows the system architecture of EarMeter. EarMeter takes two-channel in-ear audio as input and forwards it to the preprocessing module (Section 4.2.1). Here, the in-ear audio is filtered into breathing sounds and heartbeat sounds based on their distinct frequency bands, which will be processed simultaneously in separate branches. Specifically, the breathing sounds are passed to the Feature Extraction module (Section 4.2.2) to extract respiration embeddings. To enhance the encoder's ability to learn effective representations of breathing sounds across different intensities and noise interference, a teacher encoder using high-quality nasal audio as input is employed in the Knowledge Transfer module (Section 4.3) to guide the original encoder to capture effective

breathing embeddings. At the same time, for the heart sounds branch, the Respiratory Modulation module (Section 4.4) extracts breathing-modulated features from heartbeat sounds. The breathing embeddings and heartbeat features are subsequently concatenated and fed into a convolutional neural network (CNN) predictor to estimate the respiration volume. Simultaneously, earphone channel alignment is used to learn more similar embeddings between the left and right microphone channels and reduce variations from earbud placement, while output normalization is applied to reduce inter-user label variability (Section 4.5). Once training is completed, the teacher encoder branch can be removed, and EarMeter only requires in-ear audio as input during inference.

4.2 Backbone Pipeline

The backbone pipeline of EarMeter takes two-channel (left and right) in-ear audio as input and outputs the estimated respiration volume. This serves as the student model in the knowledge transfer framework, as detailed in Section 4.3.

4.2.1 Preprocessing. Before passing in-ear audio into any of the breathing volume estimation models or the respiratory modulation module, we divide audio signals into windows without overlap and process them as follows. We highpass filter the in-ear audio signals with a 50Hz cutoff, as heartbeat frequencies fall below this frequency [13, 43]. The components of the signal below 50Hz are used as input to extract heartbeat features (Section 4.4). We extract log Mel spectrograms on the highpass output, which has been shown to capture the nuances of breathing sounds better than raw waveform data [58]. We use a window size of 1024 with a hop size of 320, and 64 Mel bins ranging from 50 to 8 kHz, reshaping the audio window into a 64×964 time-frequency map. For two-channel in-ear audio, which produces two time-frequency maps, these are used as input to the encoder as shown in Figure 5.

4.2.2 Feature Extraction. Due to the scarcity of labeled data for our task, training a model from scratch often leads to poor generalization and suboptimal performance. To address this, we instead leverage the CLAP (Contrastive Language-Audio Pretraining) encoder [24], a powerful audio representation model that has been pretrained on a broad and diverse audio dataset, including human sounds, environmental noises, acoustic scenes, music, and sound effects. This extensive pretraining enables the CLAP encoder to learn generalized audio features, which can be effectively transferred to various downstream tasks.

However, while the CLAP encoder is pretrained on diverse audio datasets provides a strong foundation, it also introduces domain-specific knowledge that does not fully align with the nuances of respiratory sounds since it is pretrained on other types of sounds and texts [81]. Therefore, we tune the CLAP encoder on our breathing-specific data to better adapt the generated embeddings to the task at hand, which refines the model's ability to capture the unique characteristics of breathing sounds.

After this step, the CLAP encoder extracts two embedding vectors, each with 1024 dimensions, from each channel of the in-ear audio. These two vectors are then concatenated into an embedding matrix with a size of 2×1024 as shown in Figure 5.

4.2.3 CNN Predictor. To estimate breathing volume from in-ear audio, we pass the embedding matrix extracted by the CLAP encoder through two convolutional layers, which are used to efficiently capture relationships between the two channels. This is followed by a multilayer perceptron (MLP) layer, which projects the latent features to respiration volume as illustrated in Figure 6.

4.3 Knowledge Transfer

4.3.1 Estimating Breathing Volume from the Nose. As discussed in Section 3.3, breathing sounds from a microphone placed under the nose exhibit a high SNR, which we leverage to train an accurate estimator of breathing volume. This nose audio estimator mimics the network used in the backbone pipeline for in-ear audio but does not share

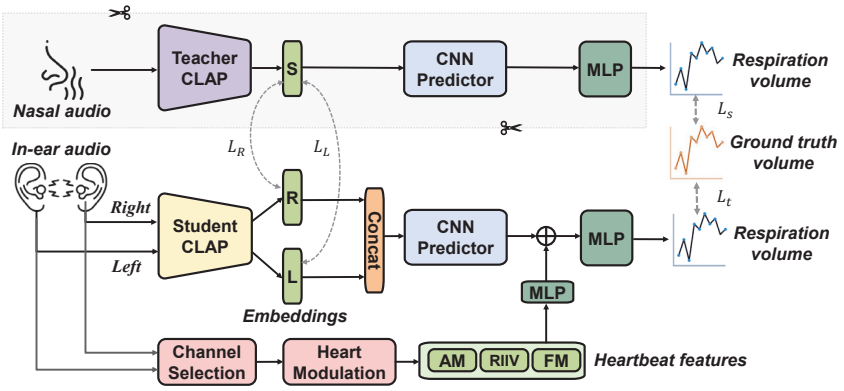


Fig. 5. Illustration of the network architecture of EarMeter. The nose audio estimator will be removed after training.

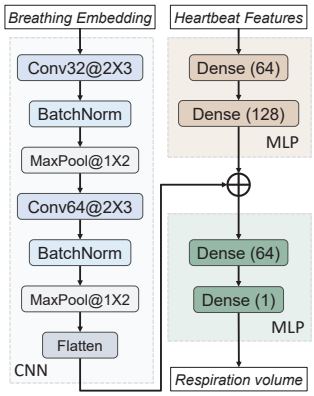


Fig. 6. Network Structures of the CNN and MLPs.

weights with it. As shown in Figure 5, to estimate breathing volume from nose audio, we first pass the nose audio through the *Feature Extractor*, and then feed its extracted embeddings into the *CNN Predictor* for breathing volume estimation. The major difference is that, instead of estimating two embedding vectors corresponding to two audio channels, the CLAP encoder of the nose audio estimator estimates a single embedding vector for the microphone placed under the nose.

4.3.2 Knowledge Transfer from Nose to In-ear Audio. Now that we have a model that can accurately estimate breathing volume from high quality recordings from a microphone under the nose, we transfer knowledge from this model to help train the student model that predicts breathing volume from much lower SNR in-ear audio data, that is often overpowered by the sounds of other activities like running or walking. Knowledge transfer, where a student model is trained to mimic the behavior of a more accurate teacher model, generally allows student models to generalize well to noisy or complex environments in a related domain [29, 32].

Architecture. Figure 5 illustrates the network architecture of EarMeter. The teacher model is the nose audio estimator built on a CLAP encoder, which is tuned using clear nasal audio signals that capture strong, uncontaminated breathing sounds. These nasal audio signals serve as an ideal reference, enabling the teacher CLAP encoder to focus on the most relevant respiratory features. The student model, based on another CLAP encoder, is trained on more challenging in-ear audio data, which may include significant interference from footstep noise and other bone-conducted sounds.

Training and Tuning. To tune the student network, we propose a novel loss function that helps guide learning across the CLAP encoders and projection layers, maintaining similarity in learned representations across the projection and encoding layers between the teacher and student models. The loss function is composed of the following components:

- *Mean Squared Error (MSE) Loss for the teacher model.* We use the MSE loss function to minimize the difference between the predicted and ground truth breathing volumes for model training. The MSE loss is defined as:

$$\mathcal{L}_t = \frac{1}{N}\sum_{i=1}^{N}(L_i - \hat{L}_i^t)^2 \tag{4}$$

where L_i represents the ground truth breathing volume for the i-th sample, and $\hat{L}_i^t$ is the predicted breathing volume from the teacher model. This loss measures the difference between the predicted breathing volume $\hat{L}^t$

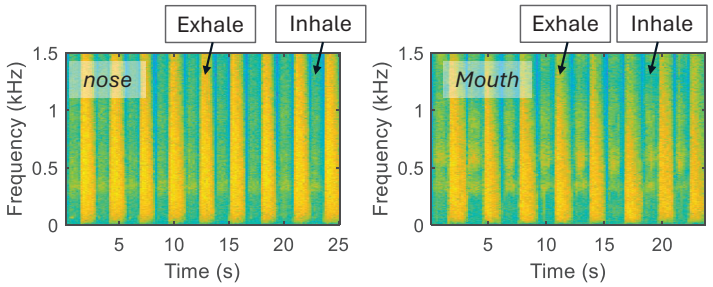


Fig. 7. Breathing sounds from nose and mouth captured by nasal microphone. We see clear breathing signatures from both modes of breathing. This allows EarMeter to incorporate training data and generalize to both modes of breathing.

using nose audio and the actual ground truth values L. It ensures that the teacher model remains accurate in predicting respiration volumes from clear nose audio.

- *MSE Loss for the student model.* Similar to the teacher's loss, this loss measures the discrepancy between the student's predicted breathing volume and the ground truth, pushing the student to make accurate predictions from the noisier in-ear audio:

$$\mathcal{L}_s = \frac{1}{N}\sum_{i=1}^{N}(L_i - \hat{L}_i^s)^2 \tag{5}$$

where $\hat{L}_i^s$ is the breathing volume predicted by the student model.

- *Similarity Loss between Teacher and Student Models.* This loss measures the Kullback-Leibler (KL) divergence [38] between the embeddings generated by the teacher and student models. The KL divergence is used to measure the difference between two probability distributions. By minimizing this loss, we guide the student model to produce respiration-related embeddings that are as close as possible to those of the teacher, ensuring that the knowledge learned from the teacher is effectively transferred to the student:

$$\mathcal{L}_{KL} = \sum_z p_t(z)\log\frac{p_t(z)}{p_s^l(z)} + \sum_z p_t(z)\log\frac{p_t(z)}{p_s^r(z)} \tag{6}$$

where z refers to the embeddings generated by the models. $p_s^l(z)$ and $p_s^r(z)$ are the probability distributions of the embeddings of the left and right in-ear audio channels, respectively. $p_t(z)$ is the probability distributions of the embeddings of the nose audio.

The total loss used for network training is a weighted average of these three components:

$$\mathcal{L}_{\text{regression}} = \alpha\mathcal{L}_t + \beta\mathcal{L}_s + \gamma\mathcal{L}_{KL} \tag{7}$$

where α, β, and γ are the constant coefficients to balance different losses. The default values of α, β, and γ are empirically set to 1 to achieve equal contributions.

Additionally, we employ an online fine-tuning strategy [41], where the teacher and student models are updated simultaneously during training. This allows both teacher and student models to be adapted simultaneously as samples are streamed in during a short data collection session (e.g., at a doctor's visit). By jointly optimizing these losses, the student CLAP model learns to accurately extract respiration-related features, even in the presence of significant interference from footstep noise. This knowledge transfer approach allows our system to maintain high performance across different breathing intensities in real-world conditions, where clear physiological signals are often difficult to isolate. After training, the teacher model is discarded, and EarMeter performs inference using only in-ear audio.

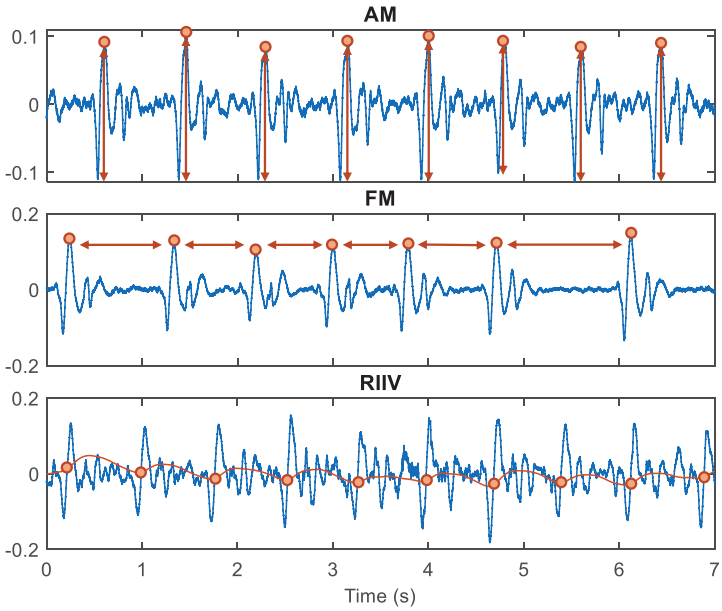


Fig. 8. Heartbeat feature extraction for respiration volume estimation.

Nose vs. Mouth Breathing. Given that people can breathe through either the nose or mouth, a natural concern is whether or not our training procedure can adapt to variations in breathing mode. Although we place the nose microphone between the nostrils and upper lip, it can still capture airflow sounds from both nose and mouth breathing, as shown in Figure 7. Moreover, during our data collection, we did not restrict users to a specific breathing mode, which allowed EarMeter to exposed to natural variations of breathing. As such, the performance of EarMeter under different breathing modes are consistent, as we show in Section 6.2.

4.4 Exploiting Heart Sounds to Boost Performance

4.4.1 Method. To further improve performance, we introduce a heart sound assisted mechanism as an opportunistic enhancement. In stationary conditions, breathing sounds are often weak, while heart sounds are usually audible and can provide complementary information. Medical studies have shown that heartbeats are modulated by the respiratory system [42], which allows heart sound dynamics to indirectly reflect respiratory effort. This coupling effect between respiration and heartbeat has been utilized in previous work for respiratory volume estimation [61]. We build on this concept and exploit the interplay between heart sounds and breathing to improve breathing volume estimation. Specifically, past works [16] have reported three key effects that the respiratory system induces on heartbeat signals, including pulse amplitude modulation (AM), pulse frequency modulation (FM), and respiratory-induced intensity variation (RIIV), as illustrated in Figure 8. We provide both qualitative and quantitative analyses to support the physiological relevance and practical utility of incorporating heartbeat-derived features to improve respiration volume estimation. The qualitative analysis draws on established physiological mechanisms underlying cardiorespiratory coupling, while the quantitative analysis examines statistical relationships between these features and breathing volume in our dataset.

Table 1. Quantitative coupling between heartbeat-derived features and breathing volume.

Metric	AM	FM	RIIV
Canonical Correlation (first component)	0.413	0.394	0.436
Mutual Information (bits)	0.062	0.041	0.078
Distance Correlation	0.127	0.119	0.133
PLSR: Variance explained (%)	12.58	10.27	13.47

1) Qualitative Analysis: We begin by reviewing physiological mechanisms that explain how respiration affects cardiovascular dynamics. These insights provide the foundational rationale for using heartbeat-derived features in our system.

a) Physiological Rationale Discussion: The physiological rationale behind incorporating AM, FM, and RIIV is grounded in well-established cardiorespiratory interactions [46] and detailed as follows:

- **Respiratory effects on pulse amplitude (AM: pulse amplitude modulation):** When we breathe in, the pressure inside our chest decreases. This affects how much blood the heart pumps out with each beat, often leading to a temporary drop in blood pressure and the strength of the pulse. As a result, the size of each heartbeat pulse, known as pulse amplitude, changes across the breathing cycle. Deeper breaths tend to cause larger changes in this amplitude, which can reflect how hard or deep someone is breathing [17]. In our system, we capture these changes using in-ear cardiovascular sounds, which reflect subtle pulse-related vibrations within the ear canal that vary with breathing.
- **Respiratory effects on pulse frequency (FM: pulse frequency modulation):** Breathing affects heart rate through autonomic nervous system control. During inspiration, heart rate typically increases; during expiration, it decreases. This pattern, known as respiratory sinus arrhythmia (RSA), is a common physiological phenomenon observed across various age groups and species [30, 80]. RSA exists as a natural mechanism to improve the efficiency of gas exchange in the lungs by adjusting blood flow in sync with breathing. The temporal fluctuation in inter-beat intervals due to RSA can serve as a proxy for respiratory depth (i.e., breathing volume). In our work, these frequency modulations are extracted from heartbeat intervals embedded in in-ear cardiovascular audio.
- **Respiratory effects on signal baseline (RIIV: respiratory-induced intensity variation):** RIIV refers to slow changes in the baseline level of cardiovascular signals that occur in sync with breathing. These changes are caused by two main factors: (1) mechanical shifts in chest pressure during breathing, which temporarily reduce the amount of blood flowing to the outer parts of the body, and (2) the nervous system adjusting the tightness of blood vessels (vascular tone) in response to each breath. When a person takes deeper or more controlled breaths, these effects become stronger and cause more noticeable shifts in signal intensity. RIIV captures these combined effects and serves as an indirect indicator of breathing effort and depth. In our system, RIIV appears as slow shifts in the in-ear audio caused by respiratory-driven changes in blood flow and vessel behavior near the ear.

b) Empirical Evidence from Related Study: In addition, recent work by Romero et al. [61] provides preliminary empirical support for the physiological rationale. Their system, OptiBreathe, uses in-ear PPG signals in a stationary and controlled setting to estimate tidal volume by analyzing respiratory-induced modulations in cardiovascular signals (including AM, FM, and RIIV). Although this study struggles to provide satisfactory performance (a MAPE of nearly 40%), it highlights the potential of heartbeat-derived features to capture meaningful information about breathing volume.

2) Quantitative Analysis: To further validate the usefulness of heartbeat features in our own dataset, we conducted a comprehensive quantitative analysis of the coupling between heartbeat-derived features and breathing volume, using multiple statistical and information-theoretic methods. These analyses assess both linear and nonlinear relationships between each feature type (i.e., AM, FM, and RIIV) and the ground truth breathing volume.
a) Analysis setup: We analyzed the statistical relationships between the extracted AM, FM, and RIIV feature sets from input audio windows and the ground-truth breathing volume using data collected under stationary conditions, where heartbeat sounds were clearly detectable. For each feature type (AM, FM, RIIV), the heartbeat-derived feature matrix had a shape of $N \times D$, where N is the number of windows and D is the number of extracted features per window. The corresponding ground-truth breathing volume was represented as an $N \times 1$ vector. Canonical correlation analysis (CCA) and Partial Least Squares regression (PLSR) were applied to the full feature matrices to evaluate multivariate relationships. For mutual information (MI) and distance correlation (dCor), we first computed a scalar summary statistic (mean) of each feature vector per window to quantify window-level dependencies with breathing volume. The full results are summarized in Table 1.
b) Interpreting the results: Our analysis demonstrates that all three heartbeat-derived feature types–AM, FM, and RIIV–exhibit statistically meaningful coupling with breathing volume as discussed below.

We first applied **Canonical Correlation Analysis (CCA)**, which measures the strength of linear association between two multivariate sets. This reveals moderate correlation values for each feature, with the first canonical component ranging from 0.394 (FM) to 0.436 (RIIV). This correlation suggests that a shared linear subspace exists between heartbeat-derived features and breathing volume.

To explore nonlinear dependencies, we computed **Mutual Information (MI)**, which captures general statistical dependence, and **Distance Correlation (dCor)**, which detects both linear and nonlinear associations. MI values ranged from 0.041 to 0.078 bits, and dCor values ranged from 0.119 to 0.133. While these values are moderate, they are consistently above zero across all feature types, confirming that each signal dimension contributes information beyond linear effects.

We also used **Partial Least Squares Regression (PLSR)**, a low-rank predictive modeling technique, to assess how well these features explain variance in breathing volume. The model explained between 10.27% (FM) and 13.47% (RIIV) of variance—modest yet meaningful results, especially considering the features were derived entirely from cardiovascular modulations embedded in in-ear audio, without using any direct respiratory signals.
3) Analysis summary. Our qualitative and quantitative analyses demonstrate that AM, FM, and RIIV carry physiologically meaningful information related to breathing volume. While the strength of coupling is moderate, the consistency across multiple statistical perspectives suggests that these features can serve as valuable supplementary signals. In our system, we incorporate these heartbeat-derived features alongside primary audio features into a learning model. Their inclusion helps capture latent cardiorespiratory dynamics and contributes to improved performance in breathing volume estimation, beyond what is achievable with respiratory audio alone.
4) Implementation details. We extract all three features (RIIV, AM, FM) from the less noisy microphone channel. To make this determination, we measure the heart rate variability (HRV) from both channels and select the channel with smaller standard deviation (STD), which we observe less likely to be impacted by noise through our deployments as lower STD implies more regular heartbeats signals.

Since heart rate varies over time, the number of values extracted for each feature will also vary between windows. To standardize the input for the model, we fill a 200-dimension feature vector with as many RIIV/AM/FM values measured within each window and pad the rest. In our deployments, we observed that maximum number of features extracted from a person is typically less than this amount, so setting the feature dimension to 200 allows us to comfortably capture the time series of every feature in a window.
5) Integrating with breathing volume estimation. To augment our breathing estimation model, we extract and assemble these features into a feature vector over an estimation window. Next, we combine this feature vector with the respiratory features extracted from the CLAP encoder and the following CNN predictor, creating

a comprehensive embedding that incorporates both direct respiratory sounds and indirect respiration-modulated heartbeat sounds, before using this newly formed embedding to estimate breathing volume. The architecture for merging these two feature spaces is shown in Figure 6.

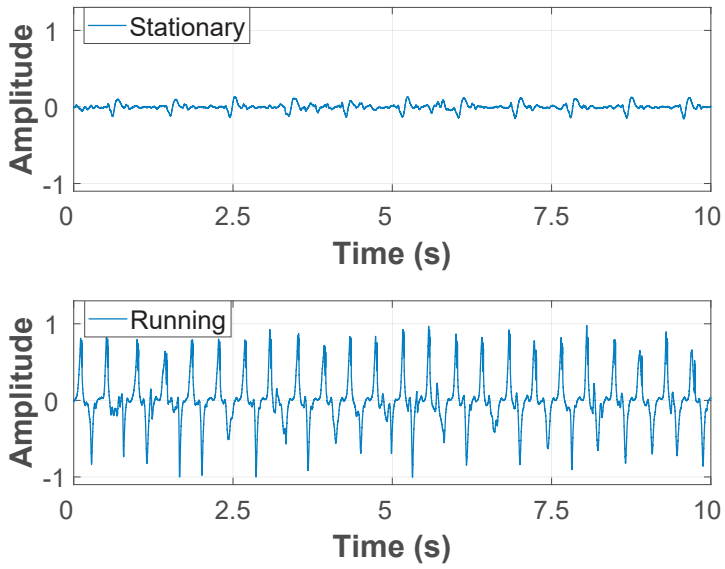


Fig. 9. Illustration of heartbeat features while standing (top) and running (bottom) from the in-ear microphone. During motion, heartbeat features are mostly overpowered. As such, we only incorporate heartbeat features when they are clearly present, which can be determined through a light-weight SVM classifier.

4.4.2 Accounting for Motion: Clean Heartbeat Detector. During high-intensity motion, such as walking or running, heartbeat sounds captured by in-ear microphones are significantly contaminated by other internal body sounds, particularly the rhythmic impacts associated with footsteps. Figure 9 shows in-ear audio collected from a user standing still (Figure 9, top) and running (Figure 9, bottom). While heart signals are discernible in the stationary condition, they become almost completely obscured during running, where the signal is dominated by strong, repetitive footstep sounds. Consequently, it becomes highly challenging to extract heart-derived features under these conditions. Therefore, **EarMeter incorporates heartbeat-derived features (AM, FM, and RIIV) only when clear heartbeat signals are detected**. Specifically, when the classifier detects clear heartbeats, these heart-derived features are extracted and appended to the extracted breathing audio features as input to the subsequent predictor for breathing volume estimation. In contrast, during high-intensity motion, when heart signals are obscured, the classifier triggers a fallback strategy that fills the heartbeat feature channels with zeros, preserving input dimensionality without introducing unreliable information.

Distinguishing heartbeat signals from high-intensity motion signals from in-ear microphones has been well explored in existing studies [43], which have demonstrated near-perfect performance. To validate this, we implemented a lightweight support vector machine (SVM) classifier operating on each incoming audio window. Motivated by the spectral and temporal profile differences observed between stationary and high-intensity conditions, we extracted a set of discriminative audio features tailored to capture key signal properties relevant for heartbeat quality assessment. Specifically, we extracted Mel-frequency cepstral coefficients (MFCCs) to capture the spectral envelope, spectral contrast to characterize differences between peaks, and root-mean-square (RMS)

energy to reflect signal amplitude variations. Features were extracted separately from the left and right in-ear channels and then concatenated to form the input representation. To optimize the model, we employed a grid search over key hyperparameters (kernel type, regularization strength, and kernel coefficient) with class balancing enabled. The final model outputs whether a given audio segment contains clean heartbeat signals or not.

The SVM is trained on 18 users' data (randomly chosen during training) and tested on the remaining 4 users to ensure user independence of the train and test sets. We implemented 5-fold cross validation and report the average results over 5 folds. The SVM classifier demonstrated strong performance, achieving an accuracy of 99.0%, precision of 99.0%, recall of 99.3%, and F1-score of 99.2%. These results highlight the model's high reliability in distinguishing between heart-clear and not-clear segments. The performance was consistent with results reported in existing studies [43].

4.5 Improving Generalizability to Different Individuals

Due to differences in physiology, the breathing sounds and patterns generated by different individuals may diverge significantly, which can impact performance. While it is possible to fine-tune the model to an individual [37], the process would involve data collection with a medical grade device, making it unwieldy to set up. Moreover, the placement of both right and left earbud channels may vary, resulting in varying levels of earbud occlusion and observed breathing or heart intensities. This introduces variability even within the same person. We introduce two additional mechanisms to improve EarMeter's ability to generalize to more people: 1) label normalization to mitigate inter-user variability and 2) earphone channel alignment to combat intra-user variability.

1) Label Normalization. We observed that the scale of ground-truth breathing volumes varies across subjects due to physiological differences, which may lead to uneven optimization during training. To mitigate this, we apply min–max normalization to the labels based on the global minimum and maximum values across all training data:

$$\tilde{y}_i = \frac{y_i - y_{\text{min}}^{(\text{train})}}{y_{\text{max}}^{(\text{train})} - y_{\text{min}}^{(\text{train})}} \tag{8}$$

where $y_{\text{min}}^{(\text{train})}$ and $y_{\text{max}}^{(\text{train})}$ are computed from all training samples, and y_i and $\tilde{y}_i$ represent the labels before and after normalization, respectively. This standardization helps stabilize the learning process and improve generalizability by reducing label scale bias across different users. During inference, we apply the inverse normalization to recover the original range of values.

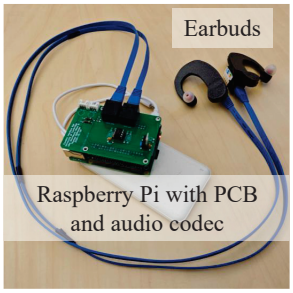
2) Earphone Channel Alignment. The signals observed by the right and left in-ear microphone channels may vary in intensity and characteristics on the same person due to variations in their placement, even if they are observing the same breathing event. To mitigate these effects, we introduce an alignment framework, we call "earphone channel alignment", that guides EarMeter to learn the same embeddings for both right and left microphones, in spite of these variations. Let f_l and f_r denote the embeddings of the left and right channels from the same user, learned from Feature Extractor as shown in Figure 4. The alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \text{sim}(f_l, f_r) \tag{9}$$

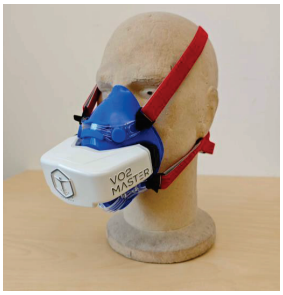
where the similarity metric we employ, $\text{sim}(\cdot, \cdot)$, is cosine similarity. This alignment loss is added to the original loss from Equation 7 to obtain:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{regression}} + \lambda \cdot \mathcal{L}_{\text{align}} \tag{10}$$

with λ controlling the trade-off between the task and alignment loss. This design encourages alignment of microphone channels to produce more general features pertaining to breathing volume that are less prone to overfitting based on confounding factors such as earbud placement.



(a) The EarMeter earable prototype.



(b) VO2Master Analyzer Mask for ground truth collection.

(c) One participant sitting at rest.

(d) One participant running on the treadmill.

Fig. 10. System implementation and experiment setting.

5 Implementation and Data Collection

5.1 Prototyping

Although commercial ANC earphones contain in-ear microphones, APIs to access the data streams remain closed to the public. As such, we designed and implemented a custom earphone prototype, as shown in Figure 10(a), which consists of 3D-printed earbuds containing microphones that face inside the ear canal. We selected the Knowles SPU1410LR5H-QB microphones [2] because of their flat frequency response between 10 Hz and 10 kHz, which covers both heartbeat and breathing sound frequencies. Data from the microphones is recorded using a Raspberry Pi 4 with an audio codec hat [1] and a custom PCB for amplification of the audio signal. We power the device using a portable power bank and place it into a chest-worn bag to ensure portability. The sampling rate of our in-ear microphone is set to 44,100 Hz.

5.2 Data collection

Ground Truth (GT) Device. We used the VO2Master Analyzer mask [4] (Figure 10(b)) for ground truth data collection. The VO2Master is an oxygen consumption (VO_2) analyzer that enables real-time monitoring of respiratory volume. To ensure accurate airflow volume measurements, the device was calibrated using 1L and 3L lab-certified air syringes under non-exercising and exercising conditions, respectively. Additionally, we attached a microphone beneath the participant's nose (positioned between the nostrils and the upper lip) to capture high-SNR nose (and/or mouth) audio. Both ground truth data and nose (and/or mouth) audio were collected solely for model training and evaluation purposes and are not required during system usage (i.e., inference).

Impact of Ground Truth Device on Breathing. Because participants need to wear the VO2Master mask to collect ground truth, there is a concern that wearing the mask may impact how users naturally breathe or the breathing sounds captured by the in-ear microphones. To assess this, we conducted a controlled empirical comparison using recordings from the same participant under identical conditions (natural breathing while stationary), with and without the mask. As shown in Figure 11, visual inspection of the spectrograms revealed no noticeable distortion or abnormal spectral artifacts introduced by the mask.

To further quantify the similarity, we computed both time-domain and frequency-domain metrics. The Mean Squared Error (MSE) between the two recordings was 0.0347. Given that the signals are normalized to the range −1 to 1, this corresponds to an average root-mean-square (RMS) amplitude difference of approximately 18.6%, indicating a relatively small deviation in waveform energy. Additionally, the Mel-Cepstral Distance (MCD), a

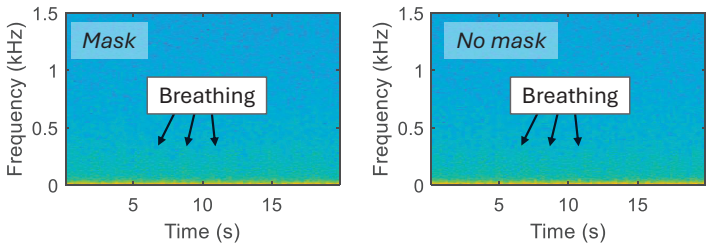


Fig. 11. In-ear audio captured with and without wearing the VO2Master ground truth collecting mask. The differences in captured signals are small, giving empirical evidence that wearing the ground truth device over the face has little impact on the captured sounds of breathing.

perceptually motivated metric widely used in audio and speech analysis, was 3.81, suggesting a moderate level of similarity in spectral envelope and perceptual characteristics.

As the recordings were made separately, small differences in breathing intensity, timing, or background conditions are expected and unavoidable. Nevertheless, the results consistently suggest that the presence of the mask does not introduce substantial distortion to the in-ear breathing signals. This aligns with our understanding that earbud microphones primarily capture internal vibrations through bone or tissue conduction, and that the air-conduction path is likely minimal or mostly blocked by the earbuds itself.

While a direct comparison of the estimated tidal volumes with and without the ground-truth mask would be desirable, such a test is inherently unreliable because a participant's breathing volume naturally varies across cycles and over time, even under controlled conditions. Instead, our analysis focuses on the acoustic consistency of the captured signals, which provides a more objective basis for assessing mask influence. Furthermore, similar respiratory measurement masks (e.g., VO2Master, COSMED K5, or metabolic carts) have been widely adopted in prior respiratory monitoring studies [6, 34, 57], establishing them as standard tools for obtaining reliable ground-truth respiration data.

Data Collection Procedures. Our experiment was approved by the Ethics Committee at our institution. When collecting data, we aimed to induce breathing at different intensities under both stationary and moving scenarios. To achieve this, we collected data while resting, exercising, and cooling down. Specifically, we recorded simultaneous in-ear audio, nose audio, and ground truth data while participants sat at rest for 10 minutes (Figure 10(c)), ran on a treadmill for 10 minutes (Figure 10(d)), and then cooled down for 10 minutes. Before running on the treadmill, participants were asked to select any two running speeds: one that was a comfortable pace for them (i.e., a jog) and one that required more effort with higher intensity (i.e., a faster run). They ran at each speed for five minutes. No specific breathing rates, intensities, or modes were imposed on participants in order to capture natural breathing patterns during different activities. As a result, the collected data naturally reflect a range of breathing volumes and intensities. In addition, to further evaluate the system's robustness across varied real-world conditions, we performed supplemental data collection in diverse acoustic environments, including outdoor settings, as well as indoor scenarios with background music and conversational speech. While these additional recordings were not part of a large-scale data collection effort, they provided valuable insight into the system's performance under more realistic and acoustically challenging conditions.

Participant Demographics. We collected data from 22 healthy participants (11 females and 11 males), totaling 660 minutes of recordings. Participants ranged in age from 23 to 53 years (mean = 29.6, SD = 6.8), with heights between 158 and 195 cm and weights from 50 to 94.3 kg. To highlight diversity, we summarize key demographic characteristics in Table 2. Based on Body Mass Index (BMI), 14 participants fell within the normal range (18.5–24.9), while two were underweight (<18.5), four overweight (25–29.9), and two obese (≥30).

Table 2. Summary of participant demographics (N=22)

Variable	Category	Count
Gender	Female	11
	Male	11
Age Group	20–29 years	12
	30–39 years	8
	40–59 years	2
BMI Category	Underweight (<18.5)	2
	Normal (18.5–24.9)	14
	Overweight (25–29.9)	4
	Obese (≥30)	2
Runner Status	Runner	7
	Non-runner	15

Additionally, seven participants self-identified as regular runners. This diverse participant pool supports the generalizability and robustness of our findings across varying body types, fitness levels, and age groups.

Respiration volume statistics. The overall distribution of the respiration volumes in the collected data is provided in Figure 12 (top), and the distribution of volumes per activity (*i.e.*, breathing intensity) is provided in Figure 12 (bottom). The dataset has an overall mean respiration volume of $1.1L$ with a standard deviation of $0.6L$. In the resting and the cooldown conditions, we see a much lower mean with a narrower standard deviation ($0.6\pm0.2L$ and $0.8\pm0.3L$ respectively), while the running condition has a much larger range and mean volume ($1.7\pm0.6L$). The difference in the respiratory volumes for each of the three conditions compared to one another is statistically significant with $\rho < 0.05$ using a Wilcoxon signed-rank test [62].

5.3 Model Training

We implement the neural network of EarMeter using PyTorch. Our Feature Extraction module incorporates the 2023 version of CLAP [24]. We optimize the network using stochastic gradient descent with a batch size of 32. During training, both in-ear audio data and simultaneously collected nose audio data are used. For inference, only the in-ear audio data is used as input. We train the model and evaluate the system's performance using leave-one-subject-out (LOSO) setting.

6 Evaluation

6.1 Evaluation Metrics

We evaluate EarMeter performance using the following evaluation metrics:

Mean Absolute Error (MAE). The average error of the prediction, defined as the average of the absolute difference between the ground truth measurement and the prediction for each window. The MAE is given in units of L (liter).

Mean Absolute Percentage Error (MAPE). The average percentage error of the prediction (given in % or as a decimal *i.e.*, %/100), computed as the average of the ratio between i) the absolute value of the difference of the ground truth measurement and the prediction, and ii) the ground truth measurement for each window.

Pearson Correlation (PC, r). The Pearson Correlation is a measure of the linear relationship between two variables (in this case the prediction and the ground truth). It measures both the strength and direction of the

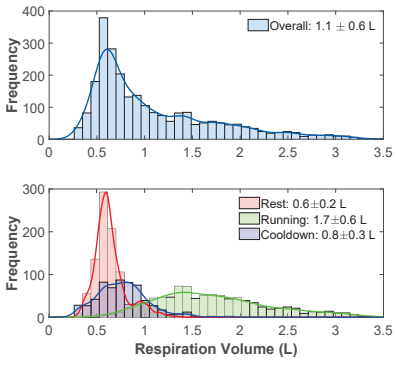


Fig. 12. Ground truth distribution.

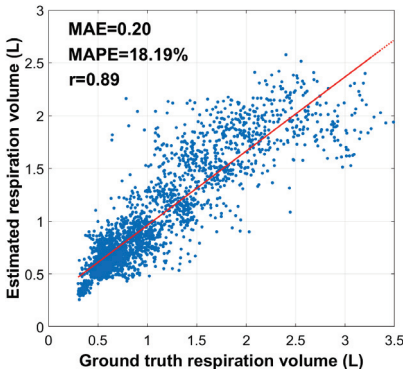


Fig. 13. Overall performance.

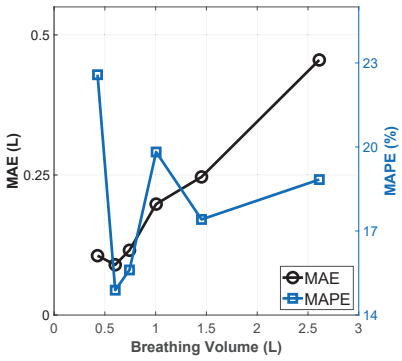


Fig. 14. Impact of breathing intensities.

relationship where $r = 1$ represents a perfect positive correlation (the desired case for this system), and $r = 0$ represents no correlation.

6.2 Overall Performance

We first evaluated the overall performance of EarMeter, followed by its performance across a series of related factors encountered in practical scenarios.

Overall Performance. As shown in Figure 13, EarMeter achieves a MAE of 0.20L and a MAPE of 18.19%, meeting the clinically required standard of 20%, under LOSO. The scatter plot in Figure 13 compares the respiration volume predictions from EarMeter with the corresponding ground truth values. The predictions and labels are well correlated, achieving a PC coefficient (r) of 0.89. The predictions fit accurately across the entire range of ground truth respiration volumes, with only a few sparse outliers. The error is symmetrically distributed around the ground truth values, indicating no significant bias, overestimation, or underestimation in the predictions.

Performance across breathing intensities. Figure 14 shows the overall performance of EarMeter across different breathing intensities. We partition the ground-truth respiration volumes into six ranges with equal numbers of samples. Each marker in the figure represents the center of a range. The overall distribution of respiration volumes spans from 0.25 L to 3.5 L, covering the intensity levels observed in humans from rest to moderate exercise [55]. Overall, EarMeter performs well across varying intensity levels. MAPE is smaller than 20% across all intensities except for the first one, which is because at lower respiration volumes, even small absolute errors contribute proportionally more to the overall percentage error due to the smaller scale of the values. At higher respiration volumes, while MAE increases slightly due to the larger scale, these larger values lead to relatively lower percentage errors. This performance pattern indicates that EarMeter maintains reliable accuracy across a wide range of intensities, making it a robust and effective system.

Performance under different scenarios. Figure 15 shows the overall performance of EarMeter under both stationary and moving scenarios. This experiment evaluates the system's performance with faint breathing sounds or in the presence of footstep interference from in-ear audio. We can observe that EarMeter delivers promising performance in both challenging scenarios, i.e., faint breathing sounds at rest and slightly stronger breathing sounds amidst footstep noise during running. It achieves a MAPE of 17.3% in stationary conditions and 19.6% in moving conditions, both well within the clinically required standard of 20% [65].

Performance under different speeds. Furthermore, we present the system's performance at different speeds, with varying breathing intensities, in Figure 16. The breathing intensity ranges between [0.18-3.00), [0.53-3.26), and [1.45-3.49) liters, corresponding to the speed ranges [4,6] km/h, (6,8] km/h, (8,12] km/h, respectively. Overall,

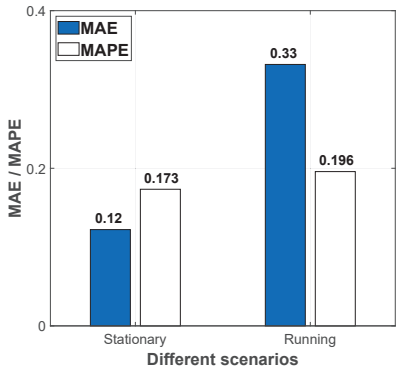


Fig. 15. Impact of scenarios.

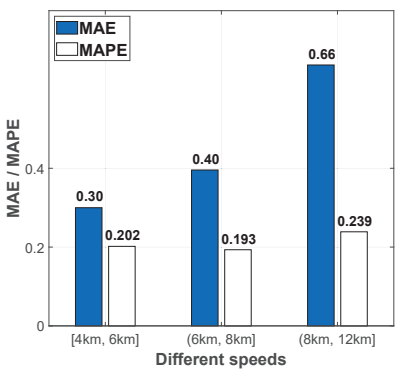


Fig. 16. Impact of different speeds.

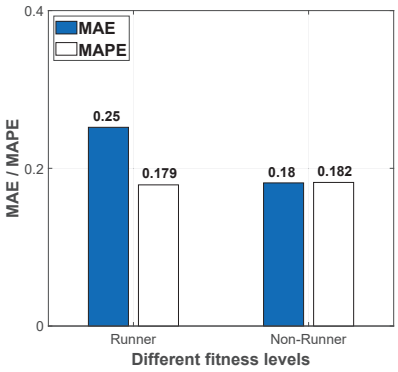


Fig. 17. Impact of fitness levels.

EarMeter achieves satisfactory performance across these conditions, with MAPE values of 20%, 19%, and 24%, respectively. The slightly higher error in the speed range of (8, 12] km/h is attributed to a subset of samples with extremely high breathing volumes, resulting in a noticeable distribution shift from the training data and lowering the average performance. Nevertheless, these results indicate that the system is robust to varying levels of footstep interference and a wide range of breathing intensities.

Impact of fitness levels. We also explored EarMeter's behavior across participants of varying fitness levels, specifically runners and non-runners, as shown in Figure 17. Runners or people of higher fitness tend to have higher average breathing volumes (in our study: 1.18 L vs. 1.03 L for non-runners), which naturally leads to a slightly higher MAE. However, the MAPE remains consistent between the two groups (17.9% for runners vs. 18.2% for non-runners), indicating that EarMeter maintains consistently high accuracy across different fitness levels.

Performance across individuals: Figure 18 reports the overall performance of EarMeter for each subject under the LOSO setting. It is evident that EarMeter achieves satisfactory performance across users, with most participants (19 out of 22) lower than 20% MAPE, meeting the clinically required standard of 20%. The higher MAPE observed for User 5 is due to this participant's relatively low breathing volume compared to other participants (i.e., a significant distribution shift compared to the training data). To better understand this case, we further examined the Pearson correlation coefficient of User 5, which remains high at $r = 0.91$. This indicates that while the model struggles with predicting the absolute scale for this user, it still accurately captures the temporal dynamics and overall trend. In the future, we expect that with a larger and more diverse dataset during the initial data collection and training phase (*e.g.*, through collaborations with sports centers, community health programs, or clinics), it will help improve the model's ability to generalize to such cases in the future.

Performance under noise levels. We evaluated the performance of EarMeter under varying noise levels in indoor and outdoor settings.

We first evaluate the performance of EarMeter under different ambient noise conditions using the overall dataset. During data collection, ambient noise levels ranged from 30 dB to 60 dB, influenced by treadmill operation and occasional background sounds such as corridor activity or air conditioning. To assess robustness, we grouped the data by noise level and evaluated model performance across these groups. As shown in Figure 19(a), EarMeter maintains consistent performance across all three noise levels, achieving near 20% MAPE in each case. This robustness is due to the occlusive design of the earbuds, which effectively block external noise and minimize its impact on the captured signals.

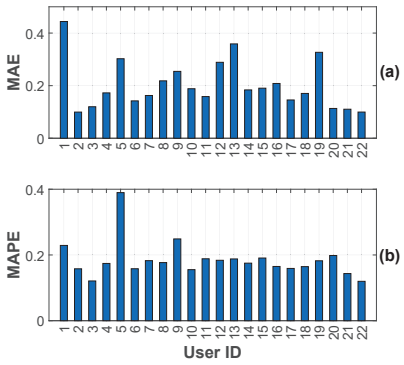


Fig. 18. Impact of individuals.

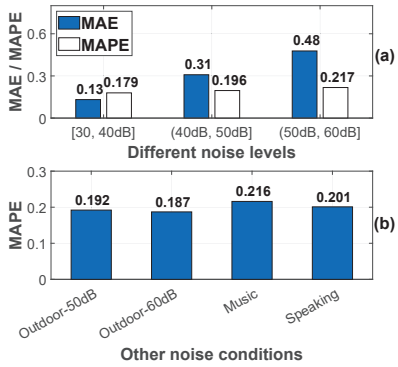


Fig. 19. Impact of noise levels.

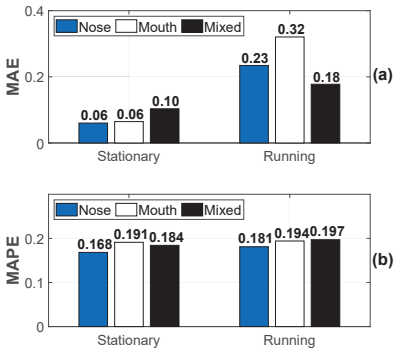


Fig. 20. Impact of breathing modes.

To further evaluate the system's robustness under more diverse acoustic conditions, we collected additional data under the following settings:

- Outdoor environments: We collected data in areas near active construction sites and traffic, where noise levels ranged from 50 to 60 dB. These settings included various ambient sounds such as vehicle engines and construction machinery.
- Indoor environments: We collected data in rooms with background music playing through a speaker and/or active conversations occurring nearby. Noise levels were maintained around 62 dB to simulate real-world indoor environments such as offices, cafés, or gyms.

Figure 19(b) presents a comparison of the system's performance across these noisy scenarios. Overall, the results demonstrate EarMeter's robustness to acoustic interference, with minimal performance degradation observed under different noise conditions. This robustness is largely attributed to the strong occlusion effect provided by in-ear acoustic sensing, which helps isolate internal physiological signals from ambient noise.

Performance under different breathing modes. In the previous experiments, participants breathed as they naturally would. To test the performance of EarMeter under different breathing modes, we conducted additional data collection where participants were asked to breathe through the nose only, mouth only, and in a mixed pattern, under both stationary and running conditions. The results are shown in Figure 20, which demonstrate that the model maintains consistent performance across all modes, with MAPE values remaining below 20%. This robustness can be attributed to two factors. First, during our original data collection, participants were not restricted to a specific breathing mode, allowing the model to be exposed to natural variations, including nose, mouth, and mixed breathing patterns, during training. Second, as discussed in Section 4.3.2, the nose microphone, positioned between the nostrils and the upper lip, is capable of capturing airflow sounds from both nose and mouth breathing for training the teacher model that guides our student model.

Power and Latency. We implement EarMeter on a Samsung Galaxy S24 smartphone and measure the latency and energy consumption of EarMeter. In summary, it requires 488.4 ms to process our input window size of 7 seconds (i.e., real-time). Through the Android Power Profiler, continuously performing inference consumes 168 uA of battery. If the phone only runs EarMeter continuously, the Samsung Galaxy S24's battery capacity of 4, 000 mAh would last for more than 990 days. In practice, smartphones typically operate for only a few days per charge, so this additional power draw has a negligible impact on overall battery life.

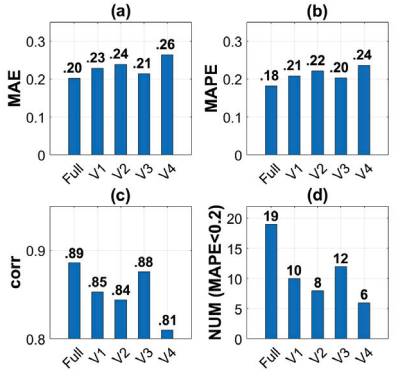


Fig. 21. Performance of ablation studies.

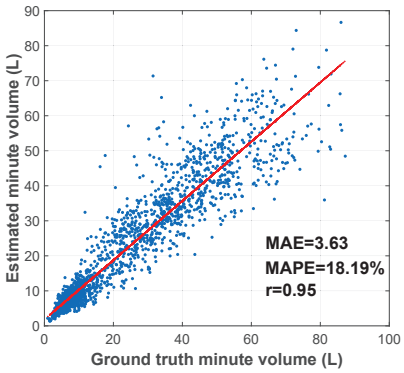


Fig. 22. Minute volume estimation.

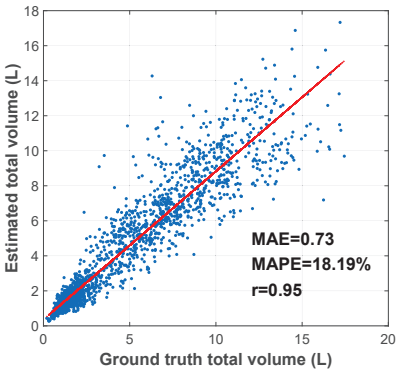


Fig. 23. Total volume estimation.

6.3 Ablation Study

Next, we present an ablation study to assess the effectiveness of each technical component proposed in EarMeter. We systematically remove each module individually, as well as all design components together, resulting in four intermediate variants and one full version of EarMeter for comparison.

- **EarMeter (Full):** The complete version of EarMeter using the CLAP backbone and all three technical components: 1) generalization design (ear channel alignment + label normalization), 2) knowledge transfer, and 3) heart feature incorporation.
- **EarMeter w/o Generalization (V1):** The full model with the generalization design component removed, i.e., without earphone channel alignment and label normalization.
- **EarMeter w/o Knowledge Transfer (V2):** The full model with the knowledge transfer component removed.
- **EarMeter w/o Heart Features (V3):** The full model with the heart feature incorporation removed.
- **EarMeter Baseline (V4):** A simplified version of EarMeter with all three technical components removed, i.e., only the CLAP backbone is used, without generalization, knowledge transfer, or heart features.

Impact of Generalization Design (V1). As discussed in Section 4.5, we incorporate label normalization and "earphone channel alignment" to improve the generalizability of our model to new participants without fine tuning, which would require specialized measurement devices. The ablation results in Figure 21 (V1) demonstrate the impact of removing this design: only 10 out of 20 subjects achieve a MAPE below 20% without these components. In contrast, incorporating these additions enables EarMeter to achieve a MAPE below 20% and meet the clinical standard for respiratory volume estimation on 19 out of 22 subjects. This highlights the effectiveness and robustness of our generalization strategy in supporting consistent performance across diverse individuals.

Impact of Knowledge Transfer (V2). As ablation results shown in Figure 21 (V2), removing the knowledge transfer module leads to a performance drop: MAE increases from 0.20 L to 0.24 L, and MAPE rises from 18.19% to 22.16%. The Pearson correlation also decreases from 0.89 to 0.84, demonstrating the effectiveness of the knowledge transfer design. In addition, 14 out of 22 subjects exceed the 20% MAPE threshold without knowledge transfer, compared to only 3 when using our complete model (Full). These results show that by leveraging high-quality nose audio to guide the student model, EarMeter is able to accurately predict breathing volume from lower-SNR in-ear audio.

Impact of Integrating Heartbeat Features (V3). Comparing the ablation results in Figure 21 (V3) with the full model (Full), it is evident that incorporating heartbeat features improves performance. This addition reduces the MAE from 0.21 L to 0.20 L and the MAPE from 20.26% to 18.19%, while the Pearson correlation increases from 0.88 to 0.89. Notably, the number of subjects with MAPE > 20% decreases from 10 to 3. These results suggest that

integrating heartbeat features provides complementary physiological information that enhances the model's ability to estimate breathing volume more accurately.

Impact of Removing All Three Design Components (Baseline) (V4). When all three technical components, i.e., generalization design, knowledge transfer, and heartbeat feature integration, are removed, the model performance drops significantly. As shown in Figure 21 (V4), the MAE increases from 0.20 L to 0.26 L, and the MAPE rises from 18.19% to 23.58%. The Pearson correlation also declines from 0.89 to 0.81. Furthermore, the number of subjects achieving MAPE below 20% drops sharply from 19 to only 6. These results highlight the substantial contribution of each design component to the overall performance and generalizability of EarMeter.

6.4 Case Studies

Minute Volume Estimation. Minute volume, defined as the total amount of air inhaled or exhaled per minute, is an important metric for assessing overall respiratory effort in daily and exercise contexts [34]. To evaluate EarMeter's performance on longer temporal scales, we compute minute volume by multiplying the model-predicted average volume per breath by the corresponding number of breathing cycles within each one-minute segment measured from the Zephyr BioHarness 3.0 chest strap [3], and compare the results with the ground truth. As shown in Figure 22, the model achieves an MAE of $3.63L$, a MAPE of 18.19%, and a correlation of 0.95. The relatively larger MAE reflects the accumulation of small per-breath estimation errors over multiple cycles, while the low MAPE and high correlation demonstrate that the model maintains accuracy and robustness when aggregating predictions over extended durations.

Total Volume Estimation. We also consider total air volume, defined as the total amount of air inhaled or exhaled within a given analysis window. This metric provides a practical view of the model's ability to capture overall respiratory effort, which is important for many long-term monitoring and activity-level analyses. The total volume is obtained by multiplying the predicted average tidal volume per window by the number of breathing cycles measured from the Zephyr BioHarness 3.0 chest strap [3]. We then compare the estimated total volume with the ground truth reference. As shown in Figure 23, the model achieves an MAE of $0.73L$, a MAPE of 18.19%, and a correlation of 0.95. The relatively larger MAE mainly results from the accumulation of small per-breath errors over multiple cycles within each window, while the MAPE and correlation remain comparable to those in average tidal volume estimation, indicating consistent performance across aggregation levels.

This case study indicates that the system performs consistently well across various breathing volume-related metrics. The comparable performance in estimating per-breath metrics (e.g., average tidal volume) and aggregate metrics (e.g., minute volume and total volume) demonstrates that our model maintains accuracy and robustness across different temporal scales. These results highlight the broad applicability and reliability of our approach.

7 Discussion

Subject variability and scaling to larger, diverse populations. Our initial study included 22 healthy participants across a wide range of ages (20-60), runners and non-runners, and equal males and females. Experimental results and ablation studies demonstrate the viability of EarMeter in adapting to new subjects, even without fine-tuning. Across participants, EarMeter achieved a MAPE below 20% for all but three individuals, thus meeting the clinical accuracy standard in most cases. For the few participants where performance degraded, the main contributing factors were distributional differences from the majority cohort and limited available training data. Nevertheless, our findings indicate that many practical challenges for real-world deployment, such as variability introduced by different breathing styles (e.g., mouth vs. nose breathing), can be largely mitigated when sufficient high-quality data are available. We expect that future studies with larger and more demographically diverse cohorts will further reduce inter-subject variability and enhance the model's generalizability.

Generalizability. Although it is possible to fine-tune EarMeter to a new user using a small amount of data from the target user, the process would involve a controlled setting where the user needs to wear specialized measuring devices (VO2Master mask). As such, it is important to consider the possibility of creating a model that can generalize to most users out-of-box. In this work, we introduced mechanisms to improve zero-shot user generalizability (label normalization and "earphone channel alignment") that enabled EarMeter to meet the clinical standard (MAPE < 20%) on all but three participant. This provides promising evidence that a model that can generalize to a wide range of participants without fine-tuning is possible. We plan to further optimize our training and inference pipelines and conduct larger studies to explore this in future work.

Clinical applicability. In this work, we explored the possibility of estimating breathing volume with earables in daily living conditions. Having visibility into breathing volume provides a picture of general health and wellness as we age and serves as an early sign of respiratory illnesses. Beyond daily living scenarios that we explored in this work, breathing volume is also an important biomarker in clinical contexts. Healthcare providers use breathing volume to identify and manage sleep disorders [33], diagnose and track respiratory diseases such as chronic obstructive pulmonary disease (COPD) [6] and asthma [57], and monitor patients under anesthesia or mechanical ventilation [23].

Although EarMeter demonstrates strong performance in estimating breathing volume under controlled conditions with healthy individuals, it is important to emphasize that the present work represents an early step toward potential clinical translation rather than a clinically validated system. All experiments were conducted in non-clinical environments, and the model has not yet been evaluated in patient populations. Further validation, calibration, and robustness analyses across broader health conditions will be necessary before EarMeter can be considered for clinical use. In future work, we plan to explore and evaluate the applicability of EarMeter in clinical settings.

8 Conclusion

In this paper, we introduce EarMeter, the first system to leverage in-ear microphones for continuous respiration volume estimation seamlessly across varying breathing intensities. Our approach provides a non-intrusive and accessible solution for monitoring respiratory volume in healthy individuals during everyday life. EarMeter leverages (1) a pretrained-model-based feature extraction technique, (2) the breathing–heartbeat coupling effect, (3) a nose-audio-based knowledge transfer strategy, and (4) a generalization strategy to address four key challenges: limited labeled data, faint breathing sounds, interference from footsteps, and generalization to unseen users. Experimental results demonstrate the effectiveness of EarMeter across diverse daily conditions. This work lays the foundation for using earphones as personal health companions for continuous and unobtrusive respiratory monitoring, paving the way for future wellness and preventive-health applications.

Acknowledgments

This research was supported by ERC project 833296 and EPSRC grants EP/Y035925/1, EP/Z53447X/1, and the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (T2EP20124-0046).

References

[1] 2023. ReSpeaker 6-Mic Circular Array Kit for Raspberry Pi | Seeed Studio Wiki. https://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/

[2] 2023. SPU1410LR5H-QB. https://www.digikey.co.uk/en/products/detail/knowles/SPU1410LR5H-QB/3621629.

[3] 2023. Zephyr BioHarness 3.0 chest strap. https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf.

[4] 2024. VO2 Master. https://vo2master.com/.

[5] 2025. Spirometry. https://www.nhs.uk/conditions/spirometry/.

[6] Ahmed M Al Rajeh and John R Hurst. 2016. Monitoring of physiological parameters to predict exacerbations of chronic obstructive pulmonary disease (COPD): a systematic review. *Journal of Clinical Medicine* 5, 12 (2016), 108.
[7] American Lung Association. 2024. Lung Capacity and Aging. https://www.lung.org/lung-health-diseases/how-lungs-work/lung-capacity-and-aging.
[8] Alessandra Angelucci, David Kuller, and Andrea Aliverti. 2020. A home telemedicine system for continuous respiratory monitoring. *IEEE Journal of Biomedical and Health Informatics* 25, 4 (2020), 1247–1256.
[9] British Lung Association. 2020. Breathing and Lung Function Tests. https://www.asthmaandlung.org.uk/sites/default/files/Section%202.1%20-%20tests%20to%20measure%20your%20breathing.pdf.
[10] Gregor Brüllmann, Karsten Fritsch, Robert Thurnheer, and Konrad E Bloch. 2009. Respiratory monitoring by inductive plethysmography in unrestrained subjects using position sensor-adjusted calibration. *Respiration* 79, 2 (2009), 112–120.
[11] Deborah Anne Burton, Keith Stokes, and George M Hall. 2004. Physiological effects of exercise. *Continuing Education in Anaesthesia, Critical Care & Pain* 4, 6 (2004), 185–188.
[12] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, Yang Liu, and Cecilia Mascolo. 2024. An evaluation of heart rate monitoring with in-ear microphones under motion. *Pervasive and Mobile Computing* 100 (2024), 101913.
[13] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. heart: Motion-resilient heart rate monitoring with in-ear microphones. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 200–209.
[14] D Caldiroli and L Minati. 2007. Early experience with remote pressure sensor respiratory plethysmography monitoring sedation in the MR scanner. *European journal of anaesthesiology* 24, 9 (2007), 761–769.
[15] Yetong Cao, Dong Ma, Wentao Xie, Qian Zhang, and Jun Luo. 2025. ESPIRO: Natural Pulmonary Function Monitoring via Earphone-Acquired Speech. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking (Hong Kong, China)(MobiCom'25). Association for Computing Machinery, New York, NY, USA*. 1–16.
[16] Peter H Charlton, Drew A Birrenkott, Timothy Bonnici, Marco AF Pimentel, Alistair EW Johnson, Jordi Alastruey, Lionel Tarassenko, Peter J Watkinson, Richard Beale, and David A Clifton. 2017. Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE reviews in biomedical engineering* 11 (2017), 2–20.
[17] Diliang Chen, Fei Chen, Alan Murray, and Dingchang Zheng. 2016. Respiratory modulation of oscillometric cuff pressure pulses and Korotkoff sounds during clinical blood pressure measurement in healthy adults. *BioMedical Engineering OnLine* 15, 1 (2016), 53.
[18] Guangwei Chen, Ildefonso de la Cruz, and Esther Rodriguez-Villegas. 2014. Automatic lung tidal volumes estimation from tracheal sounds. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 1497–1500.
[19] Michael Chu, Thao Nguyen, Vaibhav Pandey, Yongxiao Zhou, Hoang N Pham, Ronen Bar-Yoseph, Shlomit Radom-Aizik, Ramesh Jain, Dan M Cooper, and Michelle Khine. 2019. Respiration rate and volume measurements using wearable strain sensors. *NPJ digital medicine* 2, 1 (2019), 8.
[20] CP Criée, S Sorichter, HJ Smith, P Kardos, R Merget, D Heise, D Berdel, D Köhler, H Magnussen, W Marek, et al. 2011. Body plethysmography–its principles and clinical use. *Respiratory medicine* 105, 7 (2011), 959–971.
[21] Balázs Csoma, Maria Rosaria Vulpi, Silvano Dragonieri, Andrew Bentley, Timothy Felton, Zsófia Lázár, and Andras Bikov. 2022. Hypercapnia in COPD: causes, consequences, and therapy. *Journal of Clinical Medicine* 11, 11 (2022), 3180.
[22] Roberto De Fazio, Maria Rosaria Greco, Massimo De Vittorio, and Paolo Visconti. 2022. A differential inertial wearable device for breathing parameter detection: hardware and firmware development, experimental characterization. *Sensors* 22, 24 (2022), 9953.
[23] James B Eisenkraft and DL Reich. 2011. Monitoring pressure, volume, and flow in the anesthesia breathing system. *Monitoring in Anesthesia and Perioperative Care* (2011), 171.
[24] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2023. Natural Language Supervision for General-Purpose Audio Representations. arXiv:2309.05767 [cs.SD] https://arxiv.org/abs/2309.05767
[25] Biyi Fang, Nicholas D Lane, Mi Zhang, Aidan Boran, and Fahim Kawsar. 2016. BodyScan: Enabling radio-based sensing on wearable devices for contactless activity and vital sign monitoring. In *Proceedings of the 14th annual international conference on mobile systems, applications, and services*. 97–110.
[26] J Fu, W-N Teng, W Li, Y-W Chiou, D Huang, J Liu, C-K Ting, M-Y Tsou, and L Yu. 2022. Estimation of respiratory nasal pressure and flow rate signals using different respiratory sound features. *IRBM* 43, 6 (2022), 694–704.
[27] Audrey G Gift, Trellis Moore, and Karen Soeken. 1992. Relaxation to reduce dyspnea and anxiety in COPD patients. *Nursing Research* 41, 4 (1992), 242–246.
[28] Robert W Gotshall. 2002. Exercise-induced bronchoconstriction. *Drugs* 62, 12 (2002), 1725–1739.
[29] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.
[30] Junichiro Hayano and Fumihiko Yasuma. 2003. Hypothesis: respiratory sinus arrhythmia is an intrinsic resting function of cardiopulmonary system. *Cardiovascular research* 58, 1 (2003), 1–9.
[31] Christian Heyde, Hubert Mahler, Kai Roecker, and Albert Gollhofer. 2015. A wearable respiratory monitoring device–the between-days variability of calibration. *International journal of sports medicine* 36, 01 (2015), 29–34.

[32] G Hinton. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).
[33] T Hoffmann, B Eilebrecht, and S Leonhardt. 2010. Respiratory monitoring system on the basis of capacitive textile force sensors. *IEEE sensors journal* 11, 5 (2010), 1112–1119.
[34] INGVAR HOLMéR, Kalev Kuklane, and Chuansi Gao. 2007. Minute volumes and inspiratory flow rates during exhaustive treadmill walking using respirators. *The Annals of Occupational Hygiene* 51, 3 (2007), 327–335.
[35] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.
[36] Changshuo Hu, Qiang Yang, Yang Liu, Tobias Röddiger, Kayla-Jade Butkow, Mathias Ciliberto, Adam Luke Pullin, Jake Stuchbury-Wass, Mahbub Hassan, Cecilia Mascolo, et al. 2025. A Survey of Earable Technology: Trends, Tools, and the Road Ahead. *arXiv preprint arXiv:2506.05720* (2025).
[37] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.
[38] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
[39] Bernhard Laufer, Sabine Krueger-Ziolek, Paul D Docherty, Fabian Hoeflinger, Leonhard Reindl, and Knut Moeller. 2020. An Alternative Way to Measure Tidal Volumes. In *European Medical and Biological Engineering Conference*. Springer, 66–72.
[40] Jesús Lázaro, Natasa Reljin, Raquel Bailón, Eduardo Gil, Yeonsik Noh, Pablo Laguna, and Ki H Chon. 2020. Electrocardiogram derived respiration for tracking changes in tidal volume from a wearable armband. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 596–599.
[41] Lujun Li and Zhe Jin. 2022. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *Advances in Neural Information Processing Systems* 35 (2022), 635–649.
[42] Haipeng Liu, Fei Chen, Vera Hartmann, Syed Ghufran Khalid, Stephen Hughes, and Dingchang Zheng. 2020. Comparison of different modulations of photoplethysmography in extracting respiratory rate: From a physiological perspective. *Physiological Measurement* 41, 9 (2020), 094001.
[43] Yang Liu, Kayla-Jade Butkow, Jake Stuchbury-Wass, Adam Pullin, Dong Ma, and Cecilia Mascolo. 2024. RespEar: Earable-Based Robust Respiratory Rate Monitoring. *arXiv preprint arXiv:2407.06901* (2024).
[44] Yang Liu, Kayla-Jade Butkow, Jake Stuchbury-Wass, Adam Pullin, Dong Ma, and Cecilia Mascolo. 2025. Respear: Earable-based robust respiratory rate monitoring. In *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 67–77.
[45] Alexis Martin and Jérémie Voix. 2017. In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring. *IEEE Transactions on Biomedical Engineering* 65, 6 (2017), 1256–1263.
[46] David J Meredith, D Clifton, Peter Charlton, J Brooks, CW Pugh, and L Tarassenko. 2012. Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. *Journal of medical engineering & technology* 36, 1 (2012), 1–7.
[47] Javier Milagro, David Hernando, Jesús Lázaro, José A Casajús, Nuria Garatachea, Eduardo Gil, and Raquel Bailón. 2019. Electrocardiogram-derived tidal volume during treadmill stress test. *IEEE Transactions on Biomedical Engineering* 67, 1 (2019), 193–202.
[48] Vito Monaco and Cesare Stefanini. 2021. Assessing the tidal volume through wearables: a scoping review. *Sensors* 21, 12 (2021), 4124.
[49] Zahra Marjan Kazem Moussavi. 2006. *Fundamentals of respiratory sounds and analysis*. Vol. 8. Morgan & Claypool Publishers.
[50] Phuc Nguyen, Shane Transue, Min-Hyung Choi, Ann C Halbower, and Tam Vu. 2016. Wikispiro: Non-contact respiration volume monitoring during sleep. In *Proceedings of the Eighth Wireless of the Students, by the Students, and for the Students Workshop*. 27–29.
[51] Phuc Nguyen, Xinyu Zhang, Ann Halbower, and Tam Vu. 2016. Continuous and fine-grained breathing volume monitoring from afar using wireless signals. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.
[52] Andrea Nicolò, Michele Girardi, Ilenia Bazzucchi, Francesco Felici, and Massimo Sacchetti. 2018. Respiratory frequency and tidal volume during exercise: differential control and unbalanced interdependence. *Physiological reports* 6, 21 (2018), e13908.
[53] Kevin Ozment and Richard G Chang. [n. d.]. Pulmonary Issues in the Athlete/Exercise Induced Bronchoconstriction. ([n. d.]).
[54] Christophe Perrin, Carolyn D'Ambrosio, Alexander White, and Nicholas S Hill. 2005. Sleep in restrictive and neuromuscular respiratory disorders. In *Seminars in respiratory and critical care medicine*, Vol. 26. Copyright© 2005 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New ..., 117–130.
[55] Joachim D Pleil, M Ariel Geer Wallace, Michael D Davis, and Christopher M Matty. 2021. The physics of human breathing: flow, timing, volume, and pressure parameters for normal, on-demand, and ventilator respiration. *Journal of breath research* 15, 4 (2021), 042002.
[56] Mark Pollock, Jairo Roa, Joshua Benditt, and Bartolome Celli. 1993. Estimation of ventilatory reserve by stair climbing: a study in patients with chronic airflow obstruction. *Chest* 104, 5 (1993), 1378–1383.
[57] Cheng-Li Que, Christof Kolmaga, Louis-Gilles Durand, Suzanne M Kelly, and Peter T Macklem. 2002. Phonospirometry for noninvasive measurement of ventilation: methodology and preliminary results. *Journal of applied physiology* 93, 4 (2002), 1515–1526.
[58] Lawrence Rabiner and Ronald Schafer. 2010. *Theory and applications of digital speech processing*. Prentice Hall Press.
[59] Yann Retory, Pauline Niedzialkowski, Carole De Picciotto, Marcel Bonay, and Michel Petitjean. 2016. New respiratory inductive plethysmography (RIP) method for evaluating ventilatory adaptation during mild physical activities. *PLoS One* 11, 3 (2016), e0151983.
[60] Bersain A Reyes, Natasa Reljin, and Ki H Chon. 2014. Tracheal sounds acquisition using smartphones. *Sensors* 14, 8 (2014), 13830–13850.

[61] Julia Romero, Andrea Ferlini, Dimitris Spathis, Ting Dang, Katayoun Farrahi, Fahim Kawsar, and Alessandro Montanari. 2024. OptiBreathe: An Earable-based PPG System for Continuous Respiration Rate, Breathing Phase, and Tidal Volume Monitoring. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 99–106.

[62] Bernard Rosner, Robert J Glynn, and Mei-Ling T Lee. 2006. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* 62, 1 (2006), 185–192.

[63] Cheryl M Salome, Gregory G King, and Norbert Berend. 2010. Physiology of obesity and effects on lung function. *Journal of applied physiology* 108, 1 (2010), 206–211.

[64] Agnese Sbrollini, Riccardo Catena, Francesco Carbonari, Alessio Bellini, Massimo Sacchetti, Laura Burattini, and Micaela Morettini. 2022. Estimation of tidal volume during exercise stress test from wearable-device measures of heart rate and breathing rate. *Applied Sciences* 12, 11 (2022), 5441.

[65] DHT Scott and GB Drummond. 2013. III. Tidal volume measurement: OK for science, but too difficult for a workstation standard? 891–895 pages.

[66] Pragya Sharma, Xiaonan Hui, Jianlin Zhou, Thomas B Conroy, and Edwin C Kan. 2020. Wearable radio-frequency sensing of respiratory rate, respiratory volume, and heart rate. *NPJ digital medicine* 3, 1 (2020), 98.

[67] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[68] Jake Stuchbury-Wass, Yang Liu, Kayla-Jade Butkow, Josh Carter, Qiang Yang, Mathias Ciliberto, Ezio Preatoni, Dong Ma, and Cecilia Mascolo. 2025. WalkEar: holistic gait monitoring using earables. In *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 99–109.

[69] Yueyuan Sui, Minghui Zhao, Junxi Xia, Xiaofan Jiang, and Stephen Xia. 2024. TraMSR: Transformer and Mamba based Practical Speech Super-Resolution for Mobile Wearables. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 1686–1688.

[70] Mohammad Tinawi. 2021. Respiratory acid-base disorders: respiratory acidosis and respiratory alkalosis. *Archives of Clinical and Biomedical Research* 5, 2 (2021), 158–168.

[71] Michael J Tipton, Abbi Harper, Julian FR Paton, and Joseph T Costello. 2017. The human ventilatory response to stress: rate or depth? *The Journal of physiology* 595, 17 (2017), 5729–5752.

[72] Jordan Waters, Jake Stuchbury-Wass, Yang Liu, Kayla-Jade Butkow, and Cecilia Mascolo. 2024. Deep-learning based segmentation of in-ear cardiac sounds. (2024).

[73] Simon J Williams. 2004. *Chronic respiratory illness*. Routledge.

[74] Wentao Xie, Qingyong Hu, Jin Zhang, and Qian Zhang. 2023. EarSpiro: Earphone-based Spirometry for Lung Function Assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–27.

[75] Wentao Xie, Chi Xu, Yanbin Gong, Yu Wang, Yuxin Liu, Jin Zhang, Qian Zhang, Zeguang Zheng, and Shifang Yang. 2024. DeepBreath: Breathing Exercise Assessment with a Depth Camera. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–26.

[76] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. BreathListener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*. 54–66.

[77] Qiang Yang, Yang Liu, Jake Stuchbury-Wass, Kayla-Jade Butkow, Dong Ma, and Cecilia Mascolo. 2024. BrushBuds: Toothbrushing Tracking Using Earphone IMUs. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 655–660.

[78] Qiang Yang, Yang Liu, Jake Stuchbury-Wass, Kayla-Jade Butkow, Emeli Panariti, Dong Ma, and Cecilia Mascolo. 2025. SmarTeeth: Augmenting Manual Toothbrushing with In-ear Microphones. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[79] Yee Leng Yap and Zahra Moussavi. 2002. Acoustic airflow estimation from tracheal sound power. In *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No. 02CH37373)*, Vol. 2. IEEE, 1073–1076.

[80] Fumihiko Yasuma and Jun-ichiro Hayano. 2004. Respiratory sinus arrhythmia: why does the heartbeat synchronize with respiratory rhythm? *Chest* 125, 2 (2004), 683–690.

[81] Wenyu Zhang, Li Shen, Wanyue Zhang, and Chuan-Sheng Foo. 2022. Few-shot adaptation of pre-trained networks for domain shift. *arXiv preprint arXiv:2205.15234* (2022).

[82] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. 2021. MoRe-Fi: Motion-robust and fine-grained respiration monitoring via deep-learning UWB radar. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*. 111–124.

[83] Bo Zhou, Alejandro Baucells Costa, and Paul Lukowicz. 2020. Accurate spirometry with integrated barometric sensors in face-worn garments. *Sensors* 20, 15 (2020), 4234.