VitalEar: An Earable Heartbeat and Respiratory Rate Monitoring System Under Aerobic Exercises

Yuzheng Zhu, Zhangxin Liang, Jie Zheng, Yongpan Zou, *Member, IEEE*,
Victor C. M. Leung, *Life Fellow, IEEE*, and Kaishun Wu, *Fellow, IEEE*

***Abstract*—Heart rate (HR) and respiratory rate (RR) are essential physiological indicators of people's physical function and exercise performance. Advancement in sensor technology has rendered earable devices with in-ear microphones feasible for vital sign monitoring. However, it is rather challenging to monitor heart rate and respiration simultaneously with a single earable device especially when a person is doing exercises. This is because intense physical activities can lead to significant noise interference which can easily obscure physiological signals. To address this challenge, this paper presents VitalEar, an exercise physiological monitoring system based on in-ear microphones, designed to estimate HR and RR while addressing complex motion interference and variability in users and activities. VitalEaremploys Empirical Wavelet Transform (EWT) to decompose heartbeats into periodic and harmonic coefficients, enhancing noise reduction in the ECG spectrogram reconstruction model. Additionally, VitalEarincorporates a DCN-LSTM-based breathing curve reconstruction model to mitigate background noise and variability in user and activity. The experiments show that VitalEarachieves an average MAE of 5.61 BPM and 2.31 RPM, MAPE of 4.16% and 10.58% for HR and RR estimation, respectively. Compared to related work, our approach offers significant advantages in robustness against intense physical activities.**

***Index Terms*—Earable device, vital sign, heartbeat monitoring, respiration monitoring.**

I. Introduction

HR AND RR are crucial physiological indicators used for medical diagnosis and for assessing exercise intensity and fatigue levels [1]. The percentage of HR relative to the maximum heart rate (HR_{max}), calculated as ($HR_{max} \approx 220 - \text{Age}$), is an effective measure of exercise intensity. For the general population, maintaining an exercise intensity within $60\% \sim 90\%$ of HR_{max} for $15 \sim 60$ minutes is ideal [1]. RR is a reliable indicator of physical fatigue and energy retention, reflecting exercise tolerance across different populations [2]. With its high sensitivity to varying fatigue states, RR is a reliable predictor of fatigue duration [3]. Furthermore, RR effectively reflects the intensity, load, and effort of physical tasks, as well as changes in exercise volume, and is utilized in various sports activities [4], [5], [6], [7]. Excessive exercise, characterized by prolonged duration or high intensity, can lead to bodily harm and even death [8]. In ostensibly healthy individuals, abnormal HR during exercise and recovery periods is associated with an increased risk of sudden cardiac death [9]. Therefore, monitoring HR and RR during exercise helps understand one's physical condition and determine if the current exercise intensity is appropriate, thus ensuring effective workouts and avoiding overexertion.

Among existing exercise physiological monitoring devices, wrist-worn smartwatches utilizing photoplethysmography (PPG) are commonly preferred. However, during physical activity, body movement may displace the PPG sensor, alter the light path, and produce motion artifacts that impact performance [10]. Furthermore, wrist-worn exercise watches cannot monitor RR. Currently, chest straps are considered the most effective exercise physiological monitoring devices [11], but they are cumbersome to wear, require close skin contact, are prone to sweat accumulation, and are difficult to clean. Consequently, researchers have explored earable devices for physiological sensing. A. Martin et al. have utilized in-ear microphones (Mic) to capture audio within the ear canal for HR and RR estimation, demonstrating the feasibility of in-ear physiological sensing [12]. T. Ahmed et al. have reported the complementary use of IMU and in-ear Mic for RR estimation in a stationary state [13]. Noise generated during exercise may significantly affect physiological sensing performance. To address this, hEARt [14], an audio denoising method, achieves motion-resistant HR estimation. Additionally, Breathpro [15] captures noise distribution using an external Mic to aid in-ear Mic denoising, thereby facilitating respiratory pattern classification during running. These studies

Received 30 December 2024; revised 29 June 2025; accepted 28 July 2025. Date of publication 1 August 2025; date of current version 3 December 2025. This work was supported in part by China NSFC under Grant 62172286 and Grant U2001207, in part by the Key Project of Department of Education of Guangdong Province under Grant 2024ZDZX1011, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011509, in part by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007, in part by the Project of DEGP under Grant 2023KCXTD042, in part by Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, in part by Guangdong "Pearl River Talent Plan" under Grant 2019JC01X235, in part by 111 Center under Grant D25008, and in part by Shenzhen Science and Technology Foundation under Grant ZDSYS20190902092853047. Recommended for acceptance by T. X. Han. *(Corresponding author: Yongpan Zou.)*

Yuzheng Zhu, Zhangxin Liang, Jie Zheng, and Yongpan Zou are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China (e-mail: zhuyuzheng2022@email.szu.edu.cn; zhangxin.liang@foxmail.com; zhengjie_0324@163.com; yongpan@szu.edu.cn).

Victor C. M. Leung is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, Guangdong 518172, China (e-mail: vleung@ieee.org).

Kaishun Wu is with the Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511453, China (e-mail: wuks@hkust-gz.edu.cn).

Digital Object Identifier 10.1109/TMC.2025.3595178

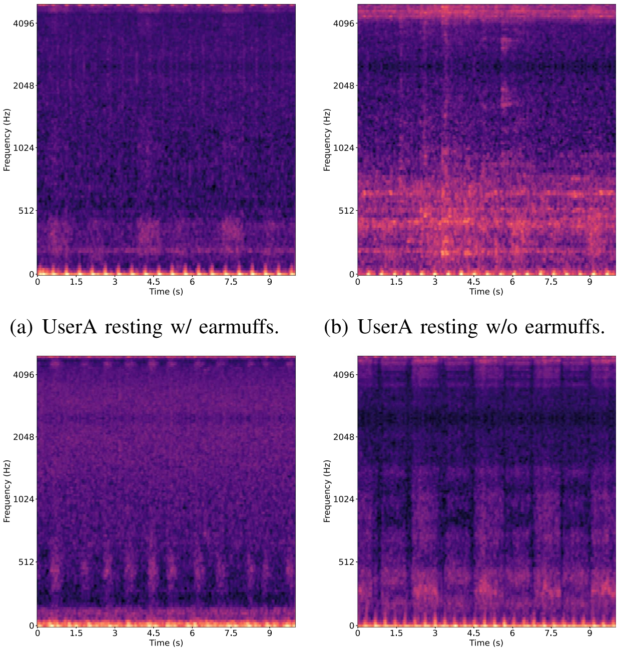
demonstrate that in-ear Mic can effectively facilitate physiological sensing in complex aerobic exercise environments.

During aerobic exercise, audio signals are significantly interfered with by body movements and background noise in the environment. Heartbeat sounds are typically distributed at frequencies below 100 Hz. Unfortunately, during exercise, friction sounds from body tissues and footsteps also fall within the low-frequency range, with intensities much stronger than those of heartbeat sounds. Consequently, heartbeat sounds are masked by noise. Furthermore, the walking pace closely matches the HR, complicating the separation of the fundamental frequency through signal processing. In terms of RR monitoring, experiments conducted in uncontrolled gym environments face challenges, as background noise shares a similar frequency distribution with breathing sounds, thereby masking them. Moreover, significant variation in the temporal and frequency distribution of breathing sounds occurs due to differences in individuals' respiratory tracts, breathing rhythms, and exercise states, challenging the robustness of the system. In Section II, we provide a detailed introduction to these challenges. To address these issues, a sports physiological parameter evaluation system, VitalEar, has been designed. For HR estimation, the empirical wavelet transform (EWT) is utilized to adaptively capture the weak period and harmonic frequencies of the heartbeat within the original audio signal. Subsequently, a ECG spectrogram reconstruction model is employed to further eliminate noise. For RR estimation, a breathing waveform reconstruction model is constructed by combining DCN and Bi-LSTM. This approach effectively captures the distribution differences of breathing sounds in both the frequency and time domains, thus reducing the impact of background noise.

The main contributions of this work are summarized as follows:

- We designed a wireless mobile earable prototype with in-ear Mic named VitalEarthat mitigates the effects of body movement, environmental noise, and variability in users and activities during exercise, and demonstrated its effectiveness in estimating HR and RR through testing with 15 subjects. To the best of our knowledge, we are the first to achieve estimation of HR and RR under complex motion conditions.
- We incorporate EWT to effectively extract more heartbeat information as additional features from low-frequency noise, and combine this with deep learning techniques to improve the accuracy of HR estimation.
- We developed a spatial and temporal feature extraction module for respiratory waveform reconstruction using DCN and Bi-LSTM. This module extracts respiratory patterns across different individuals and activities, overcoming the effects of individual differences, activity variations, and background noise, thus enabling accurate RR estimation.

The rest of this paper is organized as follows. Section II introduces the principles and noise challenges. Section III presents the system design of VitalEar. Section IV covers the implement of VitalEar. Section V presents the experimental results. Section VI discusses related works. Section VII discusses the future directions of VitalEar. Section VIII concludes our work.



(a) UserA resting w/ earmuffs. (b) UserA resting w/o earmuffs.

(c) UserB resting w/ earmuffs. (d) UserA biking w/ earmuffs.

Fig. 1. The spectrograms of acoustic signals captured in different scenarios.

II. Primariy Study

This section presents the fundamental physical principles and key challenges underlying VitalEarfor HR and RR monitoring especially under aerobic exercises.

A. *Physical Principles*

When the ear canal is blocked, low-frequency heartbeat sounds transmitted to the ear canal via bone conduction are amplified by the occlusion effect, allowing even faint heartbeat sounds to be captured [16]. For breathing, airflow passing through the narrow regions of the nasal and oral cavities creates turbulence, producing respiratory sounds that are transmitted to the in-ear microphone via bone conduction [17]. As shown in Fig. 1(d), the heartbeat signal is a short-period narrowband signal below 100 Hz, whereas the breathing signal is a long-period broadband signal ranging from 256 to 4096 Hz. The periods and frequency distributions of these signals differ significantly. By comparing Fig. 1(a) and (c), we can also observe that even under the same conditions, the heartbeat sounds of different users differ. The heartbeat sound of User A has distinct peaks, while the heartbeat sound of User B is lower and more diffuse. Additionally, the high-frequency nature of breathing sounds prevents enhancement by the occlusion effect and leads to high-frequency attenuation, resulting in much lower energy compared to heartbeat sounds. Due to the differing physical principles, significant differences exist between these two sounds.

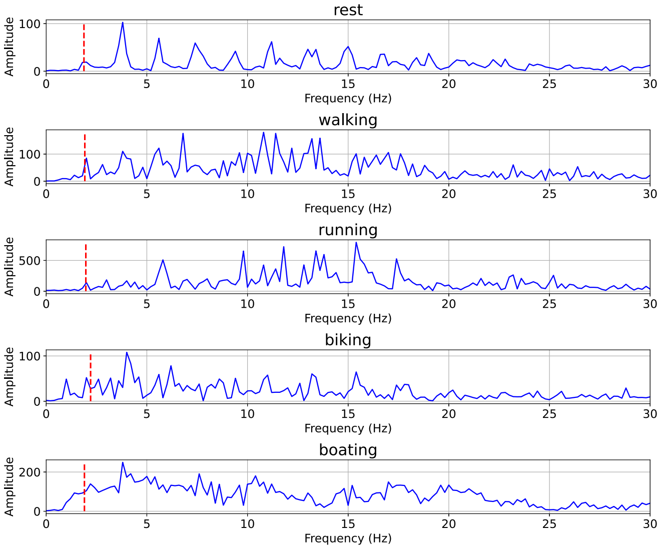


Fig. 2. The spectra of acoustic signals captured by an earphone under different activities.

B. Noise Challenge

1) Impact of Low-Frequency Noise: As shown in Fig. 2, the Fourier spectra within a 5-second window are presented for various activities (resting, walking, running, biking, boating). The red vertical line marks the average frequency of the true HR within the window. At rest, the fundamental frequency (heartbeat period) and harmonics of the heartbeat are clearly and regularly displayed in the spectrum. During walking and running, the frequency distribution of footstep sounds closely resembles that of heartbeat sounds but with higher energy, which can obscure the harmonic frequency distribution of the original heartbeat sounds. During biking, despite the absence of footstep noise, low-frequency sounds from leg and body movements—caused by friction between body tissues—are transmitted to the ear canal via bone conduction, contaminating the low-frequency harmonics of the heartbeat. In boating, the increased amplitude of body movements exacerbates this low-frequency contamination. In summary, during physical activities, friction sounds from body tissues and footsteps significantly interfere with heart sounds. This interference presents a challenge for accurately extracting the HR from the audio.

2) Impact of High-Frequency Noise: The experiment was conducted in an uncontrolled gym where numerous noise sources, such as sounds from exercise equipment and air conditioning systems, occupy similar frequency bands as respiratory sounds. As shown in Fig. 1(a) and (b), the spectrograms of audio in a resting state were compared with and without earmuffs (providing -37 dB noise reduction). With earmuffs on, background noise is blocked, facilitating the observation of breathing sound distribution. Without earmuffs, background noise masks weak breathing sounds and introduces high-energy narrowband noise, complicating the capture of effective breathing information. Additionally, the breathing frequency distribution varies among users and activities. As shown in Fig. 1(a) and (c), the breathing rhythms and respiratory tract vary among different users, resulting in User B having a breathing signal with a shorter duration and higher frequency distribution (exceeding 512 Hz). As shown in Fig. 1(a) and (d), when earmuffs are worn during biking, increased exercise intensity enhances breathing and accelerates airflow speed in the respiratory tract, resulting in a broader frequency distribution and higher intensity. These factors present challenges for accurately extracting the RR from audio.

III. System Design

A. System Overview

The system framework is illustrated in Fig. 3. Participants wear earphones equipped with an in-ear microphone to collect heart and breathing sounds from the ear canal during aerobic exercise. Simultaneously, participants wear a Zephyr chest strap [11] to obtain real ECG and breathing waveforms as ground truth (GT). During signal preprocessing, the audio signal is first decomposed into several coefficients using EWT. Lower-order detail coefficients capture low-frequency information related to heartbeat periods and harmonics, whereas higher-order detail coefficients capture information about breathing sounds. Log-Mel spectrograms of the original audio, as well as heartbeat period and harmonic coefficients, are obtained and stacked to serve as input for the ECG spectrogram reconstruction model. Simultaneously, breathing sound coefficients are pre-emphasized, and their Log-Mel spectrogram is obtained and used as input for the breathing waveform reconstruction model. For model training, the GT ECG and breathing waveform are preprocessed, and a sequence reconstruction model is used to reconstruct the breathing waveform from the audio spectrogram. Because the ECG waveform contains more details (such as the QRS complex) compared to the breathing waveform, employing a sequence model presents challenges. Consequently, a spectrogram reconstruction model is employed to reconstruct ECG spectrogram from audio spectrogram. Finally, inverse STFT processing is performed on the obtained ECG spectrogram to reconstruct the ECG waveform. HR and RR estimation are then conducted using a peak-to-peak detection algorithm.

B. Data Preprocessing

As previously mentioned, the audio signal consists of multiple sound sources. Common signal decomposition methods include Empirical Mode Decomposition (EMD) [18] and Discrete Wavelet Transform (DWT). EMD is ineffective at separating information from different modes and is prone to mode mixing; DWT employs fixed-ratio sub-band frequency division and lacks adaptive filtering capability. Therefore, Empirical Wavelet Transform (EWT) [19] is used, which adaptively selects frequency bands and overcomes mode mixing issues caused by signal time-frequency scale discontinuities.

For the signal $f(t)$, the normalized FFT spectrum of the audio is obtained, and M maxima points in the spectrum are identified and sorted in descending order. The spectrum is then divided using the minima points between each pair of adjacent maxima

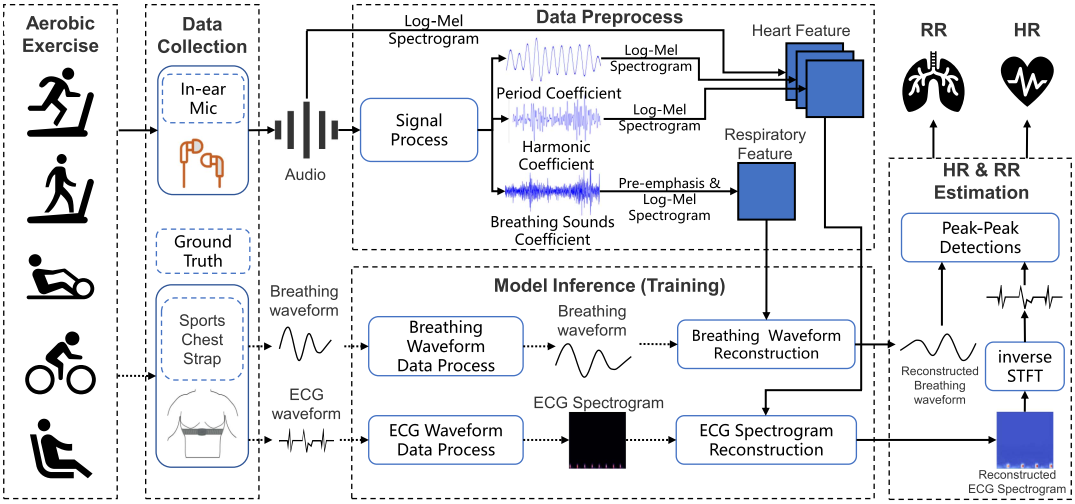


Fig. 3. The system overview of VitalEar.

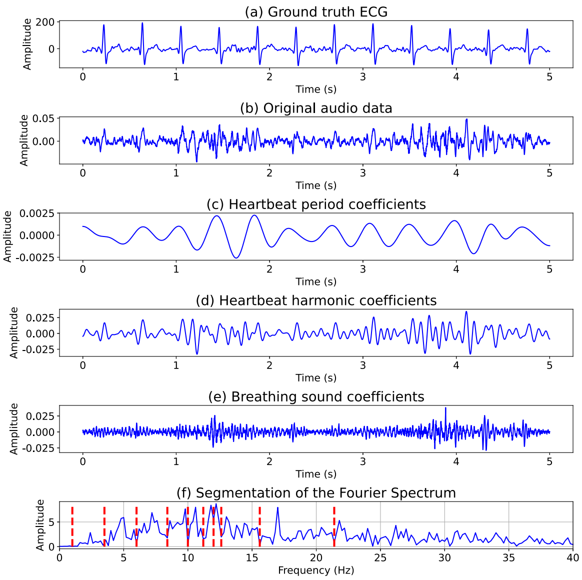


Fig. 4. The results after Empirical Wavelet Transform.

points as boundaries ω_n, where $\omega_0 = 0$ and $\omega_n = \pi$. Each segment is represented as $\Lambda_n = [\omega_{n-1}, \omega_n]$, as shown in Fig. 4(f). A transition phase T_n, centered at ω_n with a width of $2\tau_n$, is defined. The signal is decomposed into approximation and detail coefficients using EWT, with approximation coefficients typically containing only some trends of the signal, which are not useful for HR and RR estimation. Therefore, the segment of the spectrum with frequencies ranging from 0 to ω_1 is discarded to reduce the computational load. For $\forall n > 1$, given the empirical wavelet function (1):

$$\hat{\psi}_n(\omega) = \begin{cases} 1 \quad \text{if } \omega_n + \tau_n \leq |\omega| \leq \omega_{n+1} - \tau_{n+1} \\ \cos[\frac{\pi}{2}\beta(\frac{1}{2\tau_{n+1}}(|\omega| - \omega_{n+1} + \tau_{n+1}))] \\ \quad \text{if } \omega_{n+1} - \tau_{n+1} \leq |\omega| \leq \omega_{n+1} + \tau_{n+1} \\ \sin[\frac{\pi}{2}\beta(\frac{1}{2\tau_n}(|\omega| - \omega_n + \tau_n))] \\ \quad \text{if } \omega_n - \tau_n \leq |\omega| \leq \omega_n + \tau_n \\ 0 \quad \text{otherwise.} \end{cases} \tag{1}$$

The function $\beta(x) = x^4(35 - 84x + 70x^2 - 20x^3)$ is most used in [20].

We can view EWT as a filter bank composed of (1). The detail coefficients $W_f^{\varepsilon}(n,t)$ are obtained by the inner product with the empirical wavelets:

$$W_f^{\varepsilon}(n,t) = \langle f, \psi_n \rangle = (\hat{f}(\omega)\overline{\hat{\psi}_n(\omega)})^{\vee} \tag{2}$$

Therefore, the empirical mode f_k is given by:

$$f_k(t) = W_f^{\varepsilon}(k,t) \star \psi_k(t) \tag{3}$$

As shown in Fig. 4(b), some heartbeat signals in the original audio are obscured by motion noise, complicating their identification. After EWT, the corresponding coefficients are selected based on specific frequency bands. As shown in Fig. 4(c), coefficients in the $1 \sim 3$ Hz range contain the heartbeat period and resemble a sinusoidal curve, which helps determine the approximate position of the heartbeat. We add coefficients ranging from $3 \sim 20$ Hz (Fig. 4(d)), which encompass the harmonic information of the heartbeat and provide more detailed insights into ECG reconstruction. The computational complexity of searching for the maxima points in EWT increases with the sampling rate. To mitigate this, we adopt a strategy of reducing the audio sampling rate and constraining the frequency search range. Nevertheless, lowering the sampling rate may lead to the loss of high-frequency components in respiratory signals. To address this issue, the original signal is duplicated and filtered within the $256 \sim 4096$ Hz band, where respiratory components are concentrated. A wavelet-based filter is constructed to extract respiratory coefficients, followed by pre-emphasis and Mel spectrogram computation to derive respiratory features. Concurrently, the original signal is downsampled to 1 kHz and processed with EWT. To further accelerate the EWT process, the frequency search range for heartbeat signals is restricted to $0 \sim 20$ Hz. Within this range, peak detection is performed to extract the heartbeat cycle and harmonic coefficients. These features are then stacked with the original audio signal, and Mel spectrograms are computed to generate the final heartbeat

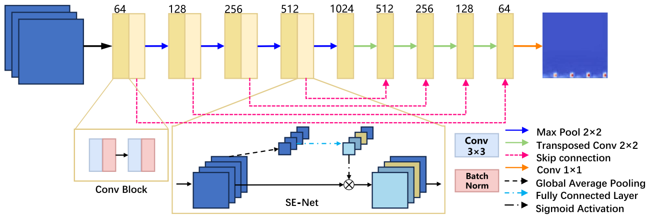


Fig. 5. The architecture of ECG spectrogram reconstruction model.

features. The window lengths for the Fast Fourier Transform are 128 for heart features and 2048 for respiratory features. The window lengths for the Short-Time Fourier Transform (STFT) are also 128 and 2048, with overlap lengths of 51 and 256, respectively. The number of Mel filters used is 64 for heart features and 128 for respiratory features. Although EWT can capture the harmonic information of the heartbeat sound, it is still affected by noise. Therefore, we employ deep learning techniques for further heartbeat extraction.

C. ECG Spectrogram Reconstruction

As shown in Fig. 5, a U-Net [21] model is employed as the base, integrated with a channel attention mechanism to effectively extract features from both the original and EWT spectrograms. In the encoder, feature maps are subjected to repeated 3×3 convolutions, LeakyReLU activation functions, and normalization layers, followed by extraction using the channel attention mechanism. SE-Net [22] is used to perform global average pooling, compressing the spatial features of each channel into a scalar. The scalar is then passed through two fully connected layers and a Sigmoid function to derive the channel attention weights. The channel attention mechanism enables the model to adaptively select effective feature channels from both the original spectrogram and EWT-derived features, thereby enhancing model performance. In the decoder, the data undergo successive up-convolution blocks, with the number of feature maps being halved at each step. After each up-convolution, the feature maps are concatenated with the corresponding encoder feature maps, followed by convolutional blocks and batch normalization. The final layer utilizes a 1×1 convolution to map the feature maps to a single output channel spectrogram, which serves as the reconstructed ECG spectrogram. During training, the GT ECG is upsampled to 1 kHz after passing through a first-order bandpass filter with a $10 \sim 50$ Hz range, and its STFT spectrogram is obtained using the same time-frequency transformation parameters as those used for heart features. We use the Adam optimizer with a learning rate of 0.002 and perform 100 iterations with a batch size of 256.

D. Breathing Waveform Reconstruction

As described in Section II, the distribution of breathing sounds varies with individual characteristics and physical activity, frequently being obscured by background noise. A deep learning framework is designed to effectively reconstruct breathing waveform from audio spectrogram, as illustrated in Fig. 6.

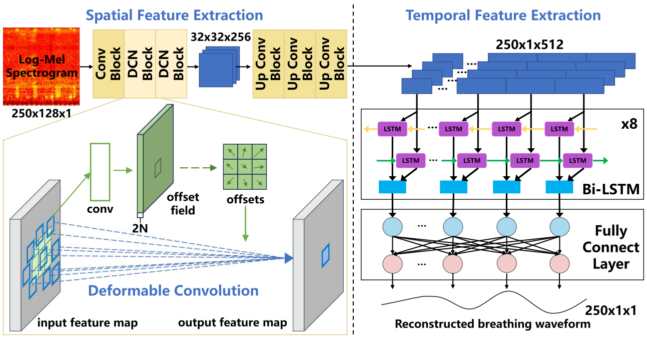


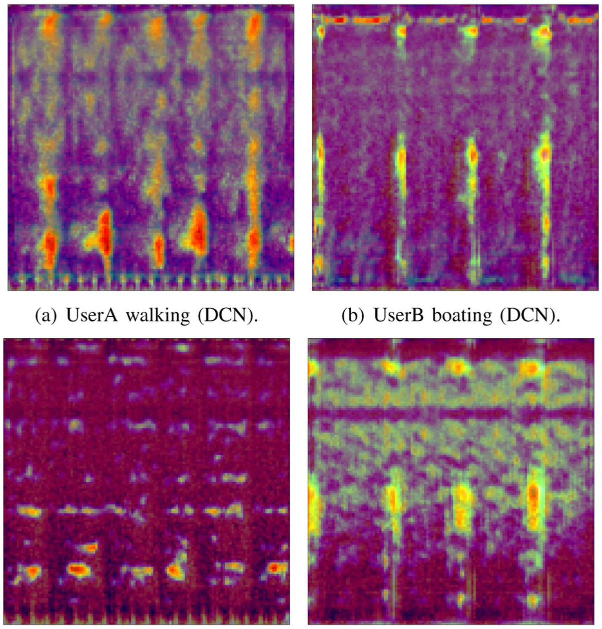
Fig. 6. The architecture of breathing waveform reconstruction model.

1) Spatial Feature Extraction: Before inputting into the model, the respiratory feature is reshaped to 250×128, where 250 represents the number of time bins and 128 represents the number of frequency bins. The window length for the breathing feature is 10 seconds, and the GT breathing waveform is sampled at 25 Hz. Reshaping the time bins to 250 ensures the temporal alignment of the reconstruction. The respiratory feature first passes through a convolutional block to extract shallow features. The convolutional block structure is the same as shown in Fig. 5. It is then processed through two layers of DCN [23] blocks to extract deep feature representations of the breathing sound. For a convolution kernel with K sampling positions, w_k and p_k denote the weight and predefined offset of the k-th position, respectively. For a 3x3 convolution, $K = 9$. $x(p)$ and $y(p)$ denote the input and output feature maps at position p, respectively. The deformable convolution formula is:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \tag{4}$$

where Δp_k denotes the learnable offset for the k-th position and is a real number with no range restrictions. Δp_k is obtained by applying a separate convolutional layer to the same input feature map, which outputs 2 K channels with kernel weights initialized to zero. The learning rate for this convolutional layer is set to 0.1 times that of the other conventional layers. After the DCN, a LeakyReLU activation function layer is applied, followed by a global pooling layer. After two DCN blocks, the output is connected to the upsampling convolution block. This upsampling convolution block structure is similar to that in Fig. 5. The key difference is that this block only expands along the time dimension while continually halving the frequency bins. Finally, it produces an intermediate feature with 1 frequency bin dimension, 512 channels, and a time length of 250. This feature is then input into the temporal information capture module.

As shown in Fig. 7, we present the Class Activation Maps (CAMs) generated by using DCN and CNN for spatial feature extraction, respectively. CAM is applied to the output of the spectral feature extractor to visualize which frequency-time regions the model attends to during heart rate and respiratory rate estimation. This enhances model interpretability and helps verify that the network is focusing on physiologically relevant features. Comparing Fig. 7(a) and (c), the feature maps from



(a) UserA walking (DCN). (b) UserB boating (DCN).

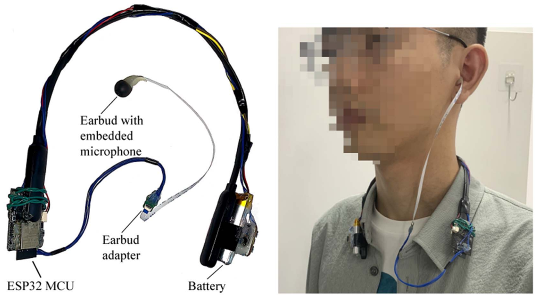
(c) UserA walking (CNN). (d) UserB boating (CNN).

Fig. 7. The CAM maps of DCN and CNN block.

DCN are more continuous and effectively capture the frequency distribution of the breathing sound. Fig. 7(a) and (b) show better robustness to frequency variations caused by different individuals and exercise intensities. In Fig. 7(b), DCN effectively isolates the breathing sound from background noise, while the feature maps from the CNN in Fig. 7(d) contain substantial background noise incorrectly identified as breathing sound. In summary, DCN demonstrates superior feature extraction capabilities for breathing sounds compared to CNN.

2) Temporal Feature Extraction: A complete breath comprises two phases: inhalation and exhalation. In the GT breathing waveform, inhalation and exhalation are represented by the rising and falling edges of the curve, respectively. In the audio time-frequency domain, these phases appear as two consecutive regions of concentrated energy. The sounds of inhalation and exhalation are very similar, rendering convolution ineffective for accurately reconstructing the breathing waveform phases. Consequently, the model inaccurately reconstructs inhalation and exhalation as the falling and rising edges of the breathing curve. A bidirectional LSTM (Bi-LSTM) is employed to capture the long-term dependencies of breathing, allowing the model to determine whether the current position corresponds to inhalation or exhalation based on contextual information. The Bi-LSTM consists of 8 layers, each with an hidden layer dimension of 256. The output is then fed into a fully connected layer to reduce the Bi-LSTM output to one dimension, resulting in the reconstructed breathing waveform.

We apply a first-order bandpass filter with a range of 0.3 to 0.8 Hz to the GT breathing waveform to eliminate motion artifacts, which is used as the ground truth for the model. We use the Adam optimizer with a learning rate of 0.002 and perform 100 iterations with a batch size of 64.



(a) Neckband earphone prototype. (b) Wearing prototype.

Fig. 8. The prototype of VitalEar.

E. HR and RR Estimation

After model inference, the reconstructed ECG spectrogram and breathing waveform are obtained. Due to the lack of phase information in the reconstructed ECG spectrogram, the Griffin-Lim algorithm [24] is employed to convert the frequency domain representation back to the time domain. A sliding window with a length of 10 seconds and a step size of 5 seconds is applied to segment the reconstructed ECG and breathing curves. Peak-to-peak detection is used for each window to calculate the time intervals between peaks. The z-score of these intervals is then computed, and outliers are filtered out. The HR and RR for each window are determined by averaging the remaining time intervals. Finally, a moving average filter is applied to smooth the HR and RR curves.

IV. Implementation

A. The Prototype of VitalEar

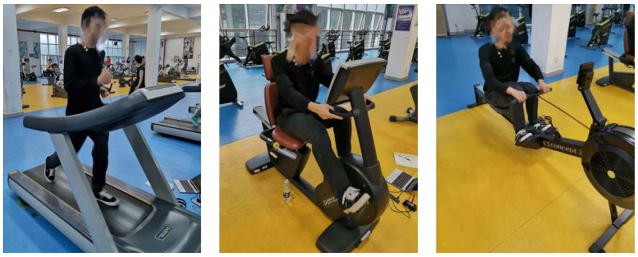
Although in-ear microphones are integrated into existing commercial earbuds (e.g., AirPods Pro), API access to their microphone outputs is not available. To collect audio data from ear canal, a neckband earphone prototype was designed. Specifically, a digital omnidirectional MEMS microphone (MEMSensing MSM261S4030H0R) was embedded into a small earbud, as shown in Fig. 8(a). The microphone is connected to an ESP32 microcontroller unit (MCU) via a flexible flat cable. The MCU acquires the audio data collected by the microphone through the Inter-IC Sound (IIS) digital audio transmission protocol. Digital transmission significantly reduces noise generated during the transmission process. The MCU transmits the audio to a laptop via Wi-Fi. During use, the device is worn around the neck, as shown in Fig. 8(b).

B. Data Collection

We involve data collection from 15 participants, comprising 5 females and 10 males, with ages ranging from 20 to 27. The participants exhibit varying levels of physical activity: 3 individuals exercise less than once per week, 5 exercise once weekly, another 5 engage in physical activity 2 $\sim$ 3 times per week, and 2 individuals exercise more than four times weekly. As shown in Fig. 9, we collect experimental data using exercise equipment

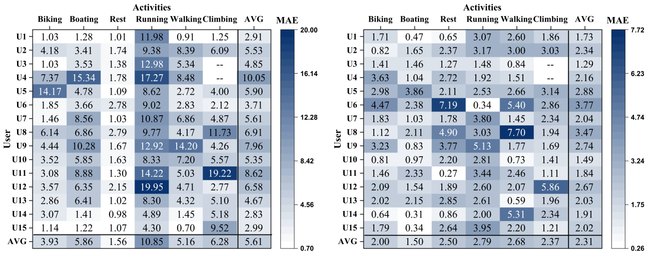
TABLE I
Ground Truth Distribution of HR and RR

GT	Biking	Boating	Resting	Running	Walking	Climbing	AVG
HR (BPM)	126.14 ± 12.23	120.02 ± 11.25	102.39 ± 12.17	131.72 ± 14.02	115.19 ± 10.91	131.70 ± 10.41	120.83 ± 15.54
RR (RPM)	25.63 ± 5.94	25.81 ± 2.81	22.77 ± 5.93	20.79 ± 4.76	20.65 ± 5.67	22.33 ± 4.74	23.02 ± 5.39



(a) Walking & running on a treadmill. (b) Biking & resting on a bicycle machine. (c) Boating on a rowing machine.

Fig. 9. The experimental setup.



(a) HR estimation. (b) RR estimation.

Fig. 10. The individual study of HR and RR estimation.

in an uncontrolled gym environment. We collect some common aerobic activities: walking, running, biking, resting, boating and stair climbing. Walking and running data are recorded on a treadmill at speeds of 3 km/h and 6 km/h, respectively, while the other activities are performed on a stationary bike and rowing machine. Each activity is sampled twice, with each sample lasting 3 minutes. Prior to data collection, users are instructed to remain still for 20 seconds to capture relatively pure heart sound data, which helps align the audio with the GT. We calculate the cross-correlation coefficient between the GT ECG and the 20-second quiet period to align the data. As shown in the Table I, we report baseline HR and RR for each participant across various activities. Among these, running and stair climbing elicit the highest average heart rate, whereas cycling and rowing yield the highest average respiratory rate. Zephyr [11] monitoring devices capture the real ECG and breathing waveform at sampling rates of 250 Hz and 25 Hz, respectively, while the prototype device samples at 10 kHz.

V. Evaluation

We use leave-one-user-out cross-validation to partition the data, where the data of one user is held out for testing, and the model is trained using data from other users. For the cross activities evaluation, We use leave-one-activity-out cross-validation to partition the data, where the data of one activity is held out for testing, and the model is trained using data from other activities. We evaluate the system's performance using the following metrics:

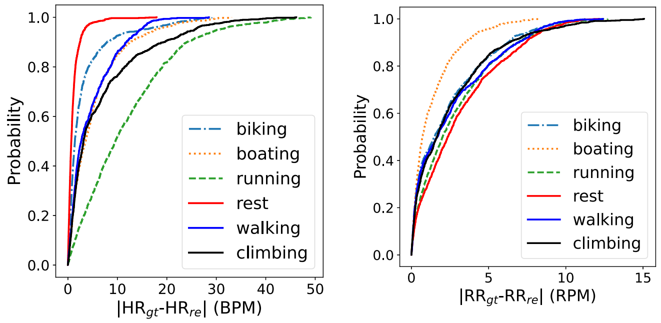
- *Mean Absolute Error (MAE):* it is the average absolute error between the groundtruth and estimation within each window.
- *Mean Absolute Percentage Error (MAPE):* it is calculated by the absolute error divided by the groundtruth within each window.
- *Bland-Altman (BA) Plots [25]:* it illustrates the range within ±1.96 standard deviations of the mean difference between two methods with 95% data points falling within this range. This metric is used for assessing the consistency between ground truth and VitalEar.

A. Overall Performance

As shown in Fig. 10, the MAE for HR and RR estimation across different users and activities is presented. Data for User3 and User4 during stair climbing have been excluded due to equipment placement issues. The following findings can be derived from the results. First, there is no correlation between HR and RR estimation. Users U5, U6, and U8 show poorer performance in RR estimation, but perform well in HR estimation. In contrast, users U4, U9, U11, and U12 show poorer performance in HR estimation, but perform well in RR estimation. This discrepancy is mainly caused by different physical properties of heartbeat and breathing. The occlusion effect significantly amplifies low-frequency heartbeat sounds, while having a minor impact on high-frequency breathing sounds. Therefore, reduced occlusion caused by improper ear seal has more severe impact on heartbeat sounds and results in higher MAE for HR estimation, while the performance of RR estimation remains stable. Second, even the same activity has different impact on HR and RR estimation. VitalEarprovides excellent HR estimation in resting, with an average MAE of 1.56 BPM, but results in poor RR estimation. This is due to reduced breathing intensity during rest, which makes breaths less distinguishable from background noise. Conversely, VitalEaryields a good RR estimation with an average MAE of 1.50 RPM but a poor HR estimation in boating. This is due to the synchronization of breathing with the boating movements stabilizes RR, while body noise from boating obscures low-frequency heartbeats, as detailed in Section II, impairing HR measurement. We also find that the boating performance of users U4, U8, U9, and U10 is poor, which is due to the high intensity of their activity during this exercise, causing excessive noise that masks the heartbeat sounds. When running, VitalEarshows the highest MAE for both HR and RR due to strong footstep noise that masks heartbeats.

TABLE II
THE BASELINE COMPARISON FOR HR ESTIMATION

	Metrics	Biking	Boating	Resting	Running	Walking	Climbing	AVG
DWT	MAE (BPM)	9.93 ± 6.90	10.96 ± 6.07	12.99 ± 19.66	15.83 ± 6.54	12.90 ± 11.47	15.40 ± 5.45	13.00 ± 11.91
	MAPE (%)	8.65 ± 6.78	9.87 ± 6.04	15.38 ± 25.35	12.91 ± 6.56	12.28 ± 12.84	12.18 ± 4.00	11.88 ± 15.60
EWT	MAE (BPM)	11.11 ± 4.71	13.60 ± 7.14	11.19 ± 9.02	15.23 ± 8.45	12.46 ± 8.74	16.80 ± 7.20	13.40 ± 10.49
	MAPE (%)	9.77 ± 5.12	12.31 ± 7.14	12.25 ± 11.55	11.88 ± 5.77	11.62 ± 9.41	12.96 ± 4.54	11.80 ± 14.76
HT	MAE (BPM)	11.57 ± 9.33	21.34 ± 11.88	24.73 ± 31.87	17.46 ± 8.80	24.73 ± 17.46	17.64 ± 6.23	19.58 ± 18.30
	MAPE (%)	16.69 ± 17.21	19.85 ± 13.05	27.86 ± 42.39	14.64 ± 8.50	26.70 ± 19.13	15.18 ± 6.04	20.15 ± 22.88
U-Net	MAE (BPM)	3.93 ± 2.96	8.49 ± 5.21	1.71 ± 0.57	12.18 ± 6.08	5.41 ± 3.99	6.75 ± 3.98	6.41 ± 5.30
	MAPE (%)	2.97 ± 2.42	7.14 ± 4.62	0.93 ± 0.39	9.29 ± 5.28	4.85 ± 4.32	5.13 ± 3.14	5.05 ± 4.56
BF-UNet	MAE (BPM)	5.06 ± 3.56	10.62 ± 5.64	1.70 ± 0.63	15.43 ± 7.45	8.06 ± 5.35	10.27 ± 5.98	8.52 ± 6.73
	MAPE (%)	4.25 ± 3.62	9.00 ± 5.09	0.93 ± 0.47	11.40 ± 4.41	7.27 ± 5.93	7.70 ± 4.18	6.76 ± 5.40
DWT-UNet	MAE (BPM)	4.45 ± 4.17	7.41 ± 5.02	1.69 ± 0.94	12.43 ± 5.03	5.23 ± 4.20	8.60 ± 5.24	6.64 ± 5.45
	MAPE (%)	3.31 ± 3.12	6.37 ± 4.89	0.92 ± 0.85	9.41 ± 3.80	4.57 ± 4.62	6.27 ± 3.79	5.14 ± 4.57
Ours	MAE (BPM)	$\mathbf{3.93 \pm 3.38}$	$\mathbf{5.86 \pm 3.85}$	$\mathbf{1.56 \pm 0.61}$	$\mathbf{10.85 \pm 4.21}$	$\mathbf{5.16 \pm 3.54}$	$\mathbf{6.28 \pm 4.81}$	$\mathbf{5.61 \pm 4.53}$
	MAPE (%)	$\mathbf{2.68 \pm 2.93}$	$\mathbf{4.74 \pm 3.08}$	$\mathbf{0.74 \pm 0.37}$	$\mathbf{8.12 \pm 2.83}$	$\mathbf{4.02 \pm 3.87}$	$\mathbf{4.67 \pm 3.44}$	$\mathbf{4.16 \pm 3.67}$



(a) ECDF (HR).

(b) ECDF (RR).

Fig. 11. The empirical cumulative distribution function of error for HR & RR estimation.

Moreover, body movements in running cause vibrations of the hardware, inducing additional interference that obscures the breathing sounds.

We have also calculated the empirical cumulative distribution function (ECDF) for HR and RR estimation and obtained the results shown in Fig. 11. We can see that about 60% of the samples have an absolute error less than 12.5 BPM and 2.5 RPM for HR and RR estimation in the worst case, respectively. What is more, among the activities considered, running and boating significantly influence HR estimation, while resting and biking have the least effect. For RR estimation, boating and biking show the least impact compared with other three activities.

B. Baseline Comparison

Table II presents the baseline comparison for HR estimation. Our system is compared with DWT, EWT [19], HT [12], and U-Net, where DWT is a common signal processing method, and HT and U-Net are methods used in previous studies [12], [14]. We also compare the experimental results of replacing EWT with a Butterworth filter (BF) and DWT while keeping the subsequent deep learning model unchanged, to demonstrate the effectiveness of EWT in adaptive filtering. VitalEarachieves the best performance in both MAE and MAPE for HR estimation. Compared to using only U-Net, the average MAE and MAPE decrease by 12.48% and 17.62%, respectively. In addition, VitalEaroutperforms the BF-UNet and DWT-UNet in all activities. This demonstrates the effectiveness of using EWT-extracted heartbeat periodic and harmonic information to provide prior knowledge for model inference during the signal preprocessing stage.

Table III presents the baseline comparison for RR estimation. Few studies address RR monitoring in dynamic conditions. Several methods are compared: HT, used for RR estimation in a stationary state as reported in [12]; and CNN-LSTM, which substitutes the DCN block with a standard CNN block for spatial feature extraction while keeping other architecture of the breathing waveform reconstruction model unchanged. The system outperforms the others in terms of MAE and MAPE. Compared to CNN-LSTM, the system reduces average MAE and MAPE by 15.38% and 20.40%, respectively, demonstrating the effectiveness of DCN in capturing the frequency distribution of breathing sounds.

C. Ablation Study

We separate spatial and temporal feature extraction in the breathing waveform reconstruction model while maintaining the same structure, obtaining two models for comparison: one using only DCN and one using only Bi-LSTM. The DCN model maps intermediate features to the breathing waveform using a fully connected layer. The Bi-LSTM model treats time bins of the input spectrogram as the temporal dimension and frequency bins as the feature dimension for each time point. As shown in Table IV, the average MAE for the DCN and Bi-LSTM models is 3.23 and 5.02 RPM, respectively, with average MAPE of 16.61% and 24.78% . This indicates that spatial or temporal feature extraction alone cannot effectively capture breathing characteristics. Combining both methods enhances the capture

TABLE III
The baseline Comparison for RR Estimation

	Metrics	Biking	Boating	Resting	Running	Walking	Climbing	AVG
HT	MAE (BPM)	3.83 ± 1.98	2.26 ± 1.76	4.58 ± 2.17	5.05 ± 2.19	4.46 ± 2.62	4.53 ± 2.58	4.12 ± 2.30
	MAPE (%)	28.34 ± 20.43	14.89 ± 13.41	32.69 ± 17.74	39.93 ± 23.09	30.49 ± 21.91	32.18 ± 24.92	29.75 ± 20.73
CNN-LSTM	MAE (BPM)	2.64 ± 1.49	1.60 ± 0.80	3.17 ± 1.62	3.07 ± 2.05	3.50 ± 3.00	2.40 ± 1.12	2.73 ± 1.61
	MAPE (%)	12.58 ± 8.70	5.95 ± 3.16	15.65 ± 10.24	15.78 ± 13.21	19.64 ± 20.27	10.16 ± 5.32	13.29 ± 9.16
Ours	MAE (BPM)	**2.00 ± 1.12**	**1.50 ± 0.96**	**2.50 ± 1.76**	**2.79 ± 1.13**	**2.68 ± 1.99**	**2.37 ± 1.24**	**2.31 ± 1.45**
	MAPE (%)	**10.35 ± 9.81**	**5.82 ± 4.48**	**9.42 ± 6.69**	**15.69 ± 12.74**	**12.10 ± 10.58**	**10.11 ± 5.36**	**10.58 ± 9.12**

TABLE IV
The Ablation Experiments for RR Estimation

	Metrics	Biking	Boating	Resting	Running	Walking	Climbing	AVG
Bi-LSTM	MAE (BPM)	4.73 ± 3.12	2.45 ± 1.21	6.62 ± 3.34	5.36 ± 2.91	5.32 ± 3.34	5.66 ± 4.51	5.02 ± 3.37
	MAPE (%)	22.57 ± 19.24	9.47 ± 4.86	32.85 ± 19.98	26.07 ± 16.58	26.63 ± 19.60	31.09 ± 32.81	24.78 ± 21.28
DCN	MAE (BPM)	3.05 ± 2.04	1.73 ± 1.24	3.66 ± 2.39	3.98 ± 2.64	3.72 ± 3.02	3.27 ± 1.89	3.23 ± 2.38
	MAPE (%)	15.78 ± 13.46	6.82 ± 5.10	18.73 ± 14.65	20.85 ± 16.35	21.29 ± 21.44	16.16 ± 11.85	16.61 ± 15.33
Ours	MAE (BPM)	**2.00 ± 1.12**	**1.50 ± 0.96**	**2.50 ± 1.76**	**2.79 ± 1.13**	**2.68 ± 1.99**	**2.37 ± 1.24**	**2.31 ± 1.45**
	MAPE (%)	**10.35 ± 9.81**	**5.82 ± 4.48**	**9.42 ± 6.69**	**15.69 ± 12.74**	**12.10 ± 10.58**	**10.11 ± 5.36**	**10.58 ± 9.12**

TABLE V
The Cross Activities Evaluation for HR Estimation

	Metrics	Biking	Boating	Resting	Running	Walking	Climbing	AVG
U-Net	MAE (BPM)	4.88 ± 4.03	9.12 ± 5.17	1.59 ± 0.56	14.88 ± 6.32	8.82 ± 6.09	7.45 ± 4.52	7.79 ± 6.28
	MAPE (%)	3.58 ± 2.80	7.68 ± 4.55	0.71 ± 0.35	11.64 ± 5.89	8.06 ± 6.69	5.85 ± 3.62	6.26 ± 5.63
BF -UNet	MAE (BPM)	5.23 ± 4.06	11.59 ± 6.84	1.53 ± 0.58	13.96 ± 5.43	7.90 ± 4.42	8.69 ± 5.35	8.15 ± 6.26
	MAPE (%)	4.13 ± 3.51	9.28 ± 4.37	0.64 ± 0.30	10.91 ± 5.16	6.39 ± 3.80	7.76 ± 5.59	6.52 ± 5.31
DWT -UNet	MAE (BPM)	4.64 ± 4.65	9.36 ± 5.65	1.52 ± 0.55	13.72 ± 5.38	8.73 ± 4.82	8.29 ± 5.91	7.71 ± 5.87
	MAPE (%)	3.55 ± 3.51	7.95 ± 4.69	0.60 ± 0.18	10.68 ± 4.96	7.81 ± 5.18	6.32 ± 4.41	6.15 ± 5.06
Ours	MAE (BPM)	**4.28 ± 4.29**	**8.78 ± 5.18**	**1.50 ± 0.61**	**13.64 ± 5.67**	**7.58 ± 4.70**	**6.92 ± 4.37**	**7.12 ± 5.77**
	MAPE (%)	**2.93 ± 3.40**	**7.17 ± 4.34**	**0.49 ± 0.21**	**10.11 ± 4.76**	**6.72 ± 5.27**	**5.36 ± 3.27**	**5.46 ± 4.94**

of both frequency and temporal aspects of breathing sounds, thereby reducing RR estimation errors.

D. Cross Activities Evaluation

This section evaluates the system's performance under cross-activity conditions. A leave-one-activity-out strategy is adopted, where data from the target activity are used for testing and data from the remaining activities are used for training. Since methods based on HT, DWT, and EWT rely primarily on signal processing, their cross-activity performance is comparable to that of cross-user evaluation and is therefore omitted here for brevity.

As shown in Table V, our system achieves a 7.65% reduction in average MAE and a 10.69% reduction in MAPE compared to the DWT-UNet baseline, demonstrating improved generalizability across activities. However, all methods—including ours—exhibited decreased HR estimation accuracy in the cross-activity setting compared to the cross-user setting. This performance degradation is likely due to the presence of activity-specific noise components not observed during training. Although signal processing techniques such as EWT help extract preliminary heartbeat features, residual noise—particularly low-frequency interference—still affects ECG reconstruction and HR estimation when the model encounters unfamiliar motion artifacts.

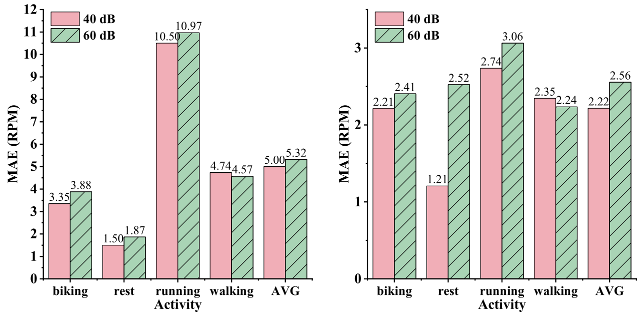
As shown in Table VI, for RR estimation, a similar trend of reduced accuracy is observed in the cross-activity setting. Nonetheless, our system consistently outperforms baseline methods. The most notable performance drop occurs in the running activity. This is due to a combination of elevated respiratory demand and biomechanical disruptions caused by rhythmic body movements. Specifically, the mechanical coupling between limb motion and respiratory musculature (e.g., diaphragm and intercostal muscles) results in non-uniform airflow and fragmented respiratory sounds. These spectrally discontinuous patterns hinder effective waveform reconstruction and degrade estimation accuracy.

E. Impact of Ambient Noise

We further evaluate the impact of ambient noise levels on system performance. As shown in Fig. 12, experiments are conducted in a gym (60 dB) and an office (40 dB). Due to equipment constraints, only four activities—biking, resting, running, and

TABLE VI
THE CROSS ACTIVITIES EVALUATION FOR RR ESTIMATION

	Metrics	Biking	Boating	Resting	Running	Walking	Climbing	AVG
CNN -LSTM	MAE (RPM) MAPE (%)	1.86 ± 1.01 8.40 ± 5.54	2.68 ± 1.40 11.02 ± 6.77	6.75 ± 2.35 32.73 ± 14.97	5.63 ± 3.37 30.36 ± 21.86	4.28 ± 3.46 23.49 ± 22.98	1.90 ± 0.93 8.37 ± 5.26	3.85 ± 2.99 19.06 ± 18.13
Ours	MAE (RPM) MAPE (%)	**1.82 ± 1.07** **8.21 ± 5.37**	**1.76 ± 0.98** **6.66 ± 4.03**	**2.53 ± 1.75** **12.54 ± 10.84**	**4.94 ± 3.44** **26.06 ± 21.72**	**3.96 ± 2.96** **22.63 ± 21.71**	**1.86 ± 0.61** **8.10 ± 3.56**	**2.81 ± 2.35** **14.04 ± 15.22**



(a) Impact of different noise on HR estimation. (b) Impact of different noise on RR estimation.

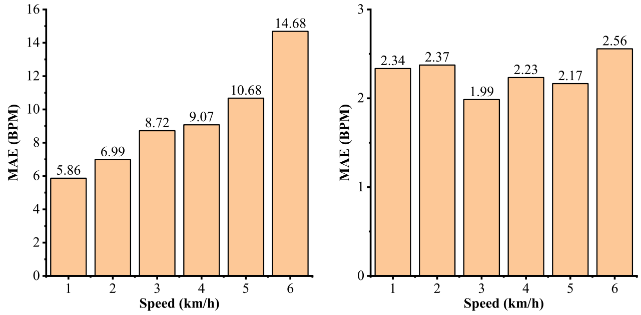
Fig. 12. Impact of different noise on HR and RR estimation.

walking—are evaluated in the office setting. Fig. 12(a) shows that HR estimation improves in quieter environments, with average MAEs of 5.00 BPM at 40 dB and 5.32 BPM at 60 dB. For RR estimation, Fig. 12(b) indicates that the resting condition benefits most from lower noise, with MAEs of 1.21 RPM and 2.52 RPM at 40 dB and 60 dB, respectively. This may be due to spectral overlap between ambient noise and breathing sounds; lower noise allows the model to capture RR features more accurately and reduce errors. Across all evaluated activities, average MAEs for RR estimation are 2.21 RPM at 40 dB and 2.55 RPM at 60 dB.

For boating and stair climbing, which require specific equipment not available in the office, participants perform these tasks in the gym while wearing noise-reduction earmuffs rated at −37 dB. Results show that for boating, the MAEs of HR and RR estimation are 9.79 BPM and 6.38 BPM without earmuffs, and 2.06 RPM and 1.88 RPM with earmuffs, respectively. For stair climbing, the corresponding MAEs are 7.97 BPM and 4.85 BPM, and 2.51 RPM and 2.19 RPM. Surprisingly, wearing earmuffs does not always improve accuracy; in some cases, it degrades performance. This may be caused by the stethoscope effect from friction between earmuffs and device cables during movement, which contaminates the audio signal and affects model inference.

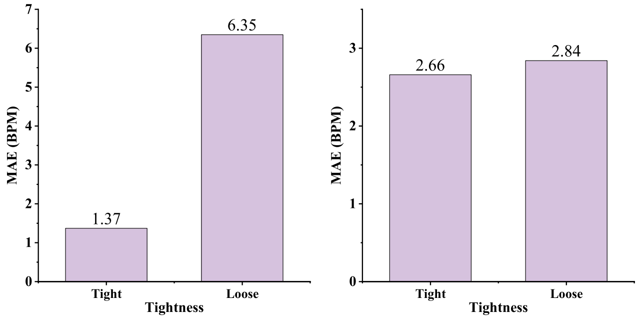
F. Impact of Different Speed

To evaluate the generalizability of VitalEarunder varying physical intensities, we conducte experiments in which users walked or jogged on a treadmill at different speeds ranging from 1 to 6 km/h. This setup simulates a range of realistic aerobic activities. As shown in Fig. 13, increasing movement speed leads to stronger body-induced noise—such as tissue friction and footstep impacts—which significantly raises the MAE in HR estimation. This trend suggests that higher locomotor intensity introduces stronger low-frequency interference, making it more difficult to extract heart sound components accurately. In contrast, RR estimation remains relatively stable across different speeds. This robustness can be attributed to the fact that respiratory sounds predominantly occupy higher-frequency bands, while motion-induced noise is primarily confined to lower frequencies. The limited spectral overlap reduces interference, thereby maintaining consistent estimation performance.



(a) Impact of different speeds on HR estimation. (b) Impact of different speeds on RR estimation.

Fig. 13. Impact of different speeds on HR and RR estimation.



(a) Impact of earplugs tightness on HR estimation. (b) Impact of earplugs tightness on RR estimation.

Fig. 14. Impact of earplugs tightness on HR and RR estimation.

G. Impact of Earplugs Tightness

We evaluate the impact of earplug fit tightness on HR and RR estimation accuracy. Since the proposed hardware system supports interchangeable earplugs of various sizes, each participant wears two different sizes under stationary conditions to simulate varying levels of fit tightness. As shown in Fig. 14(a), loose earplugs significantly degrade HR estimation, with the MAE increasing to 6.35 BPM. This is mainly due to reduced

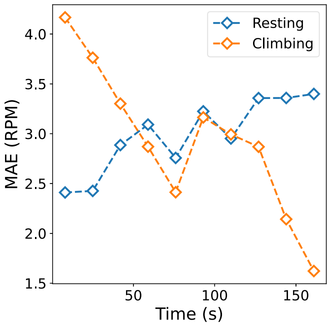


Fig. 15. The impact of respiratory intensity on RR estimation.

occlusion in the ear canal, which weakens the amplification of low-frequency heart sounds and impairs signal quality. In contrast, the impact of earplug fit on RR estimation is relatively minor, as shown in Fig. 14(b). This is because respiratory sounds mainly occupy higher frequency bands, which are less affected by occlusion loss. As a result, the system maintains stable RR estimates even under loose-fitting conditions.

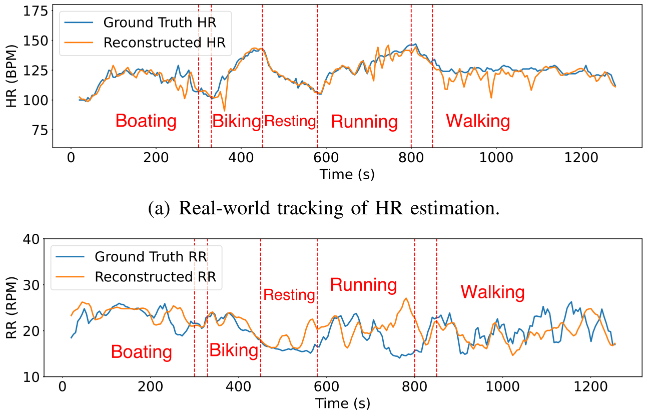
To address this issue, we can design an algorithm to determine whether the earplug is tightly. For example, by calculating the signal-to-noise ratio, we assess whether the quality of the heartbeat signal meets the required standard. If not, the system prompts the user to adjust or replace the earplugs.

H. Impact of Respiratory Intensity

Since respiratory intensity is difficult to measure directly, we assess its effect on RR estimation performance indirectly by analyzing the temporal trends in estimation error under two contrasting activity conditions: stair climbing and resting. These activities were chosen to control for confounding noise sources such as footstep or equipment noise. As shown in the Fig. 15, we plot the RR estimation error over time for all users under both conditions. During stair climbing, sustained exertion leads to progressively intensified breathing, which enhances the prominence of respiratory acoustic features. This results in a gradual reduction in estimation error over time. In contrast, during the resting phase, as users recover from exertion, their respiratory intensity declines, leading to weaker respiratory sound features and an increase in estimation error. These findings suggest that the system's RR estimation accuracy is positively correlated with respiratory intensity. In particular, higher-intensity breathing produces clearer acoustic signals, which facilitate more accurate estimation. This highlights the system's suitability for use in moderate to vigorous aerobic activities, where respiratory patterns are more pronounced and consistent.

I. Real-World Tracking: A Case Study

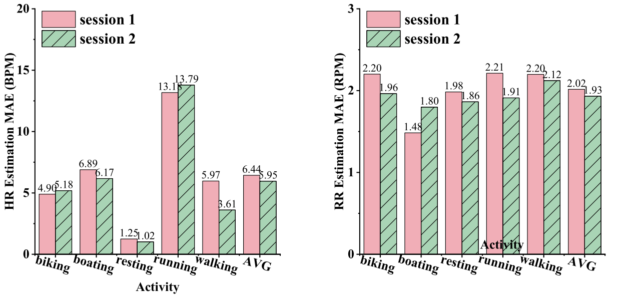
We evaluate the real-world tracking performance of users during various aerobic exercises in real-world conditions. Users engage in boating, biking, running, and walking in the gym, with rest periods between activities based on their exercise conditions. Fig. 16(a) and (b) show the real-world tracking of HR and RR, respectively. The average MAE for HR and RR estimates is 3.46 BPM and 2.40 RPM, respectively, with MAPE values of 2.78% and 12.89% . The HR estimation MAE for each activity phase is 4.00, 3.47, 3.93, and 3.44 BPM, while the RR estimation MAE is 1.35, 0.88, 4.24, and 2.50 RPM. The HR estimation performs well. Although some fluctuations are observed in RR estimation, the overall trend is favorable.



(a) Real-world tracking of HR estimation.

(b) Real-world tracking of RR estimation.

Fig. 16. The Real-world tracking performance of VitalEar.



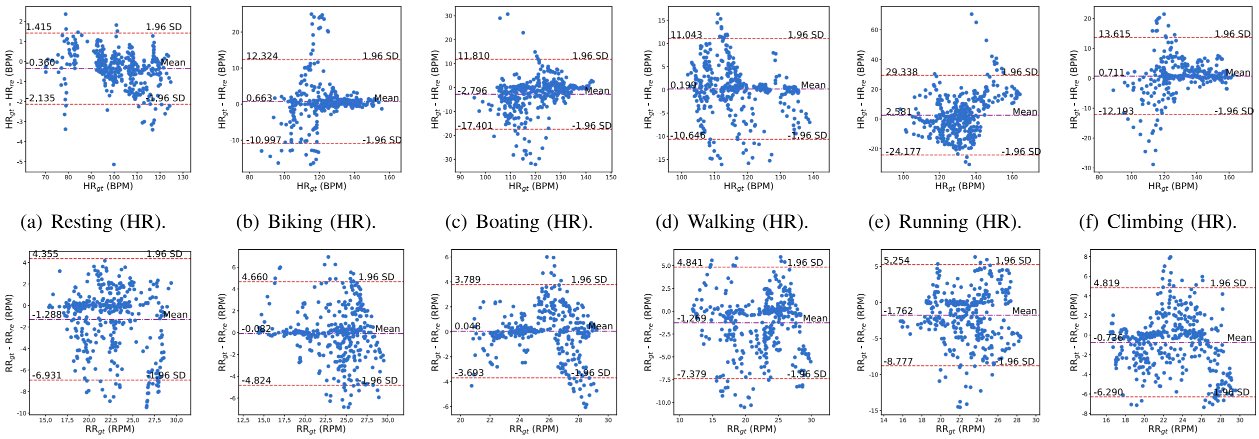
(a) Long-term performance of HR estimation.

(b) Long-term performance of RR estimation.

Fig. 17. The long-term performance of HR and RR estimation.

J. Long-Term Performance

We perform long-term experiments with VitalEarto verify the system's performance under long-term use. Four volunteers are recalled to collect new data. The time gap between data collection in Session 1 and Session 2 is approximately six months. As shown in Fig. 17, we analyze the data from Session 1 and Session 2 based on different activities. For HR estimation, the MAE difference for walking between Session 1 and Session 2 is significant, with a difference of 2.36 BPM. However, overall, the average MAE for Session 1 and Session 2 shows only a slight difference, with values of 6.43 BPM and 5.95 BPM, respectively. For RR estimation, the MAE difference for boating between Session 1 and Session 2 is relatively large, with a difference of 0.31 RPM. The MAE for RR estimation is 2.02 RPM in Session 1 and 1.93 RPM in Session 2. Session 2 outperforms Session 1, which we hypothesize is due to the data for Session 1 being collected in summer when the gym's air conditioning and fans are running, leading to higher environmental noise. In contrast,



(a) Resting (HR). (b) Biking (HR). (c) Boating (HR). (d) Walking (HR). (e) Running (HR). (f) Climbing (HR).

(g) Resting (RR). (h) Biking (RR). (i) Boating (RR). (j) Walking (RR). (k) Running (RR). (l) Climbing (RR).

Fig. 18. The Bland-Altman plots of HR and RR estimation.

Session 2 data was collected in winter when the air conditioning and fans were off, resulting in lower noise levels. Overall, the long-term experiments demonstrate that VitalEaris capable of sustained use.

K. Bland-Altman Plots

As shown in Fig. 18, the BA plots illustrate the agreement analysis between ground truth values and the estimated HR and RR under various conditions. In terms of HR estimation, the deviations between the proposed method and the ground truth remain consistently low across different activity states. Notably, during resting and biking, the mean measurement biases are -0.37 BPM and 0.66 BPM, respectively, indicating the absence of systematic bias in the proposed approach. When subjects are at resting (Figs. 4–6(a)), the limits of agreement (red dashed lines) in the BA plot are relatively narrow, suggesting high accuracy in HR estimation. However, during physical activities (Fig. 18(b)–(f)), the limits of agreement widen as exercise intensity increases, reflecting a rise in the standard deviation of the HR estimates. Compared with prior studies, the proposed method demonstrates superior performance under the same activity conditions. Specifically, the limits of agreement and mean bias during rest, walking, and running are (-2.14 to 1.42, -0.37), (-10.64 to 11.04, 0.20), and (-24.18 to 29.34, 2.58), respectively, which outperform those reported in hEARt [14]: (-10.53 to 9.58, -0.48), (-23.37 to 18.03, -2.67), and (-28.29 to 37.63, 4.67), and also exceed the performance of study [12] under static conditions: (-14.3 to 13.4, -0.44). During biking, walking, and stair climbing, the estimation bias tends to decrease as HR increases. This phenomenon may be attributed to prolonged activity leading to a stronger heartbeat, which makes the heart sounds more prominent and thereby improves HR estimation accuracy. In contrast, during boating and running, the benefits of intensified heartbeat are partially offset by strong motion artifacts, resulting in a less pronounced improvement in the BA plots compared to other activities.

As shown in Fig. 18(g)–(l), the proposed method also maintains low deviations between estimated and ground truth RR values across various activities. Specifically, during biking, boating, and stair climbing, the average biases are -0.08 RPM, 0.05 RPM, and -0.74 RPM, respectively, indicating minimal systematic bias. Moreover, the width of the BA plot's limits of agreement remains similar across different activity conditions, demonstrating that the system maintains stable performance under varying measurement scenarios. Compared with previous work, the proposed method achieves comparable or better performance. For instance, during rest and walking, the results are (-6.93 to 4.35, -1.29) and (-7.38 to 4.84, -1.27), respectively, which are comparable to the static condition in study [12]: (-2.62 to 7.48, 2.4), and superior to RRDetection [26] under walking conditions: (-11 to 11, -0.13). These findings further validate the reliability and adaptability of the proposed approach for RR estimation.

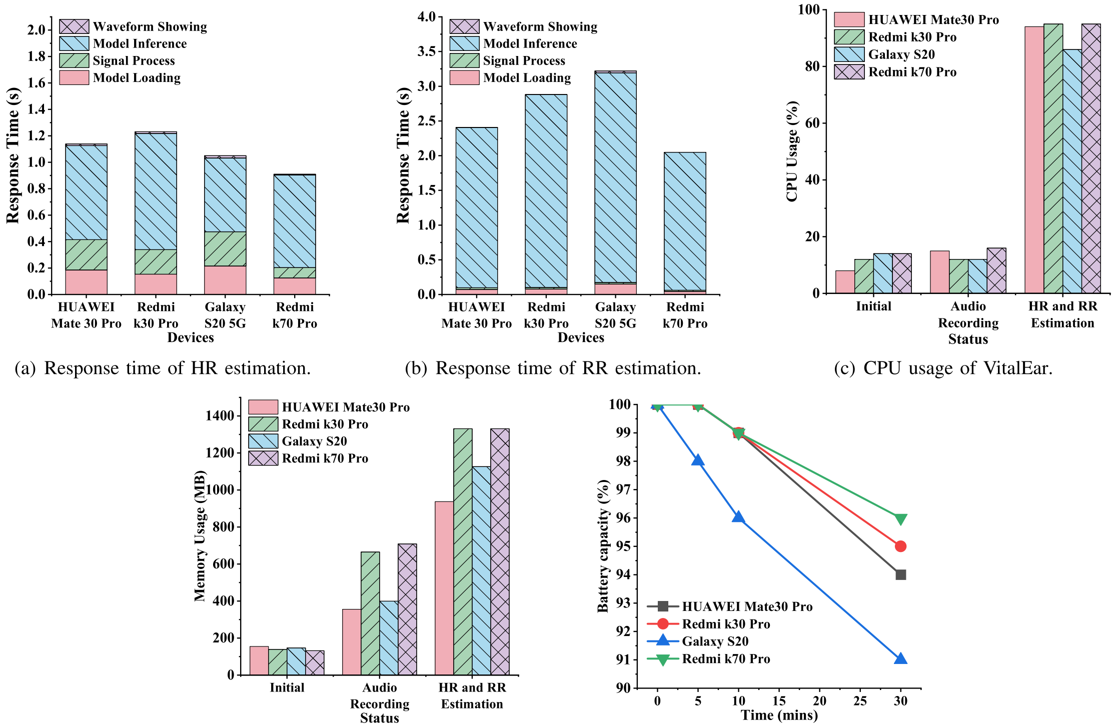
TABLE VII
DEVICE PARAMETERS

Device	CPU	RAM	Battery
HUAWEI Mate 30 Pro	2.86 GHz	8 G	4500 mhA
Redmi k30 Pro	3.05 GHz	12 G	5000 mhA
Samsung Galaxy S20	2.73 GHz	8 G	4000 mhA
Redmi k70 Pro	3.3 GHz	16 G	5000 mhA

In summary, the Bland-Altman analysis confirms the consistency of the proposed method in estimating HR and RR. The results demonstrate that the method maintains high measurement accuracy and stability across a wide range of activity scenarios.

L. Deployment Evaluation

We conduct deployment tests on the HUAWEI Mate 30 Pro, Redmi K50 Pro, Samsung Galaxy S20, and Redmi K70 Pro, with Table VII summarizing the relevant performance metrics of these smartphones. We assess the overall performance of the system when HR and RR estimation are performed simultaneously. Both the signal processing pipeline and model inference are deployed on a mobile device. Our evaluation focuses on system response time, CPU and memory usage, and power



(a) Response time of HR estimation.

(b) Response time of RR estimation.

(c) CPU usage of VitalEar.

(d) Memory usage of VitalEar

(e) Battery consumption of VitalEar

Fig. 19. The deployment examinations of HR and RR estimation.

consumption. As shown in Fig. 19(a) and (b), we analyze the response times of VitalEaracross different devices during various processing stages. When HR and RR estimation are conducted concurrently, the average inference latencies are 1.08 seconds and 2.64 seconds, respectively. Given that HR and RR typically change gradually, this level of latency is sufficient for real-time physiological feedback.

Fig. 19(c) and (d) illustrate the CPU and memory usage of VitalEaracross different processing stages on various devices. The ECG spectrum reconstruction network and the respiratory waveform reconstruction network have parameter sizes of 37.70 MB and 21.79 MB, respectively. Regarding resource usage, the models currently use 32-bit floating-point (float32) representation, resulting in relatively high CPU and memory consumption—averaging 92.5% CPU usage and 1181.6 MB memory usage. Quantization techniques can be considered in future work to reduce floating-point precision, thereby lowering memory consumption and improving computational efficiency.

Fig. 19(e) illustrates the power consumption of VitalEar. We evaluate the power usage of four devices operating continuously for 5 to 30 minutes under full battery conditions. On average, the system consumes 6% of the device's battery every 30 minutes, which is considered acceptable. Regarding the hardware evaluation, VitalEarequipped with a 1000 mAh battery. During HR and RR estimation, the system consumes approximately 0.42 W of power, corresponding to a current of 128 mA. Theoretically, this allows for up to 7.8 hours of continuous operation. However, considering real-world factors such as battery voltage fluctuations, the actual runtime is approximately 5.5 hours, which is sufficient for typical exercise monitoring scenarios.

M. Comparison With Related Work

We compare our work with other relevant studies on in-ear physiological sensing devices. Tables VIII and IX present comparisons of HR and RR estimation, respectively.

1) Activity Categories: Regarding activity categories, our study covers a variety of aerobic exercises. During these exercises, the interference noise generated is more intense than in other studies, making it more challenging to suppress noise for accurate HR and RR estimation.

2) Performance Comparison: As shown in Table VIII, for HR estimation, among passive in-ear microphone-based systems, hEARt [14] and Butkow et al. [27] uses a U-Net architecture for spectral reconstruction, while A. Martin et al. [12] apply Hilbert transform-based signal processing for HR estimation. In contrast, our method combines EWT with deep learning to extract additional heart-related features, improving ECG reconstruction accuracy and outperforming these methods across various activities. Unlike active acoustic systems such as EarMonitor [28] and APG [29], which use speaker-microphone pairs to capture detailed physiological signals, our passive system leverages bone conduction occlusion to detect heart sounds. Though active methods may perform better in high-motion scenarios like running, our approach offers comparable accuracy with lower power consumption. Compared to [12] and [30], our system offers greater interference resistance and robustness.

For RR estimation, RRDetection [26] uses the IMU sensors of commercial headphones for respiratory monitoring, while [31] combines IMU and in-ear microphones to complementarily estimate RR. However, it relies on target detection to

TABLE VIII
THE COMPARISON OF HR ESTIMATION

Literature	**Sensors**	**Metrics**	**Activities**	**Performance**
hEARt [14]	In-ear Mic	MAE MAPE	stationary walking (uncontrolled) running (uncontrolled) speaking	3.02±2.97 BPM, 4.32±3.99% 8.12±6.74 BPM, 9.53±8.28% 11.23±9.20 BPM, 9.80±7.93% 9.39±6.97 BPM, 12.06±8.88%
Butkow et al. [27]	In-ear Mic	MAE MAPE	stationary walking running	1.88±1.59 BPM, 2.11±2.61% 6.83±5.05 BPM, 7.78±6.17% 13.19±11.37 BPM, 9.60±9.28%
Earmonitor [28]	In-ear Speaker In-ear Mic	MAE	sleeping sitting speaking head moving walking (2 km/h) running (4 km/h)	1.04 BPM 1.71 BPM 3.42 BPM 4.63 BPM 6.31 BPM 7.28 BPM
APG [29]	In-ear Speaker In-ear Mic	MAPE	stationary stationary with music running active	2.64% 2.99% 3.97% 5.40%
Q. Zhang et al.[30]	Behind ear ECG & PPG	MAE	stationary	0.8±2.7 BPM
A. Martin et al.[12]	In-ear Mic	MAE	statinoary	4.3 BPM
VitalEar (ours)	In-ear Mic	MAE MAPE	resting biking boating walking (3 km/h) running (6 km/h) stair climbing	1.56±0.61 BPM, 0.74±0.37% 3.93±3.38 BPM, 2.68±2.93% 5.86±3.85 BPM, 4.74±3.08% 5.16±3.54 BPM, 4.02±3.87% 10.85±4.21 BPM, 8.12±2.83% 6.28±4.81 BPM, 4.67±3.44%

TABLE IX
THE COMPARISON OF RR ESTIMATION

Literature	**Sensors**	**Metrics**	**Activities**	**Performance**
A. Martin et al.[12]	In-ear Mic	MAE	statinoary	2.7 RPM
T. Ahmed et al.[31]	Earable IMU In-ear Mic	MAE Retention Rate	sitting, standing, lying	<2 RPM 75%
BreathPro [15]	In & Out-ear Mic	MAE	running	1.88 RPM
RRDetection [26]	Earable IMU	MAE	3-mins step test 6-mins walk test	3.87 RPM 3.75 RPM
RespEar [32]	In-ear Mic	MAE MAPE	sedentary conditions active conditions	1.48 RPM, 9.12% 2.28 RPM, 11.04%
VitalEar (ours)	In-ear Mic	MAE MAPE	resting biking boating walking (3 km/h) running (6 km/h) stair climbing	2.50±1.76 RPM, 9.42±6.69% 2.00±1.12 RPM, 10.35±9.81% 1.50±0.96 BPM, 5.82±4.48% 2.68±1.99 BPM, 12.10±10.58% 2.79±1.13 RPM, 15.69±12.74% 2.37±1.24 RPM, 10.11±5.36%

identify breathing energy clusters in discrete time-frequency windows for RR estimation, which can result in some windows failing to detect RR, leading to a lower monitoring retention rate. In contrast, our method generates breathing waveforms directly from the spectrogram, offering stronger continuity and ensuring that every window can consistently reflect respiratory frequency. Additionally, IMU signals are highly susceptible to motion interference, so [31] can only estimate RR in scenarios with minimal movement. BreathPro [15] uses both internal and external microphones for RR estimation, employing the external mic to mitigate noise interference. Our RR estimation performs slightly worse under running conditions, primarily due to the device's reliance solely on an in-ear microphone, which cannot capture external noise, making the task more challenging. RespEar [32] reports MAEs of 1.48 RPM (resting) and 2.28 RPM (walking/running), with MAPEs of 9.12% and 11.04%, respectively. Our method achieves an average MAE of 2.31 RPM and MAPE of 10.58% across all activities. While our error is slightly higher in some cases, we use a 10-second window compared to RespEar's 60-second window with 30-second overlap, offering significantly better real-time performance. Nonetheless, compared to [12], our approach can simultaneously estimate HR and RR and demonstrates greater robustness against interference.

3) Deployment Comparison: Regarding latency, hEARt [14] achieves HR estimation in 65.64 ms, and RespEar [32] requires 3.11 s and 12.27 s for its two RR estimation modes. Our system performs HR and RR estimation jointly, with respective delays of 1.08 s and 2.64 s. Although HR estimation latency is slightly longer than that of hEARt, our RR estimation is significantly faster than RespEar. Considering the slow variability of HR and RR, our system still provides near-real-time feedback suitable for practical applications.

VI. Related Work

A. Mobile Physilogical Sensing

With the advancement of smart devices, many mobile devices now have physiological sensing capabilities. Photoplethysmography (PPG) is a widely used method, extensively applied in smartwatches. Besides PPG, inertial measurement units (IMUs) are also commonly employed in physiological monitoring. For instance, D. Liaqat et al. [33] use the built-in IMU of smartwatches combined with machine learning to estimate the user's respiratory rate (RR) during daily activities. P.-Y. Hsu et al. [34] place a tri-axial accelerometer on the user's sternum to measure HR and RR before and after exercise. Additionally, Y. Cao et al. [35] extract IMU signals from commercial wristbands and use machine learning and deep learning methods to estimate electrocardiogram (ECG) waveforms. X. Guo et al. [36] design a hardware and software system based on biomagnetism to monitor HR and respiratory rate in users of different skin tones, significantly improving accuracy. However, for aerobic exercises involving high-intensity and extensive body movements, these methods may be less effective. For example, wristband devices based on PPG exhibit an average error of up to 30% during intense exercise [37]. ER-rhythm [38] use RFID tags placed on the user's limbs and chest to capture body movements and breathing patterns during exercise, estimating locomotor-respiratory coupling (LRC) through correlation methods. However, this method requires users to wear RFID tags and use antennas, limiting its practicality. Additionally, there are efforts to integrate extra sensors into masks for physiological monitoring. For example, SpiroMask [39] uses consumer-grade masks for lung function monitoring. C. Romano et al. [40] embed microphones in the mask to monitor breathing during walking and running. However, wearing a mask during aerobic exercise affects the athlete's breathing, disrupting normal performance.

B. Earable-Based Physilogical Sensing

Earable devices are easy to use, highly accepted by users, and have integrated numerous sensors, including IMUs and in-ear microphones. In recent years, researchers have made significant advancements in earable physiological sensing, including applications such as ear disease monitoring [41], identity recognition [42], [43], activity recognition [44], and more. For example, Y. Cao et al. [45] designed a general-purpose acoustic sensing platform based on commercial active noise-canceling (ANC) headphones, demonstrating its functionality and adaptability across various acoustic sensing applications. For HR estimation, methods such as [30] have utilized ECG and PPG sensors on Earable devices to monitor HR. hEARt [14], Butkow et al. [27] and [12] have leveraged the occlusion effect for HR estimation. Techniques such as Earmonitor [28] and APG [29] exploit the reflection of audio from speakers and microphones to extract HR. For RR estimation, [31] have combined IMUs with in-ear microphones to monitor RR, while BreathPro [15] uses two microphones—one inside and one outside the ear—to estimate RR during running. Most of these related works lack the capability to simultaneously estimate HR and RR and are predominantly conducted in stationary states, exhibiting poor robustness. In contrast to previous studies, VitalEarenables reliable monitoring of HR and RR under motion conditions. Specifically, we leverage EWT to preprocess audio-based heartbeat information, which enhances feature robustness for downstream deep learning models. This approach offers improved resilience to noise compared to methods that do not incorporate signal enhancement. Furthermore, we design a breathing waveform reconstruction network that integrates both temporal and spatial features, enabling the extraction of breathing signals from in-ear audio. This architecture demonstrates superior generalizability and performance in more complex physical activities and noisy environments than existing methods. A detailed comparison of related works is provided in Tables VIII and IX.

VII. Discussion

VitalEarstill has some limitations at its current stage. First, the system's real-time performance can be further improved. To reduce HR estimation errors, we uses the EWT as an audio data preprocessing method. As described in Section III, the EWT algorithm searches for M local maxima in the spectrum and separates the detail coefficients. In practical applications, as the sampling rate increases, the computational complexity of searching for these maxima also increases. In the future, we plan to optimize the performance of EWT's search algorithm, such as by exploring methods to reduce the sampling frequency or narrowing the search range, to improve operational efficiency.

Second, since different physical activities introduce distinct noise patterns, incorporating an activity classifier is indeed a promising direction. By identifying the current activity in advance, the system can apply activity-specific denoising strategies to improve the accuracy of HR and RR estimation. For instance, as shown in Fig. 2, each activity produces characteristic low-frequency noise, which may persist even after applying Empirical Wavelet Transform. In future work, we plan to design an activity recognition module and integrate it with an attention mechanism that adaptively adjusts model parameters based on the recognized activity. This approach is expected to enhance the model's robustness to activity-induced noise and further improve estimation performance.

Additionally, our prototype device has potential for further design improvements. We have already integrated an IMU sensor in the current prototype to capture chest movements with low motion interference, which helps address the issue of low respiratory intensity during the resting state, a factor that reduces RR estimation accuracy. In the future, we plan to further enhance the prototype by adding a speaker and sensors such as PPG in the earbud, enabling multi modality HR and RR estimation for more accurate results across a wider range of conditions.

As the hardware device designed in this study does not include a built-in speaker, it is not possible to conduct experiments involving simultaneous physical activity and music playback. Nevertheless, this study discusses the feasibility of such a scenario. For heart rate estimation, heart sound frequencies are primarily concentrated below 50 Hz, while typical music contains only about 1.5% of its energy in this range [46]. Therefore, the spectral overlap between heart sounds and music is minimal, suggesting that music playback has limited interference with heart rate estimation. This indicates the potential of the system to monitor heart rate during music playback. In contrast, respiratory sounds occupy the 256 Hz to 5000 Hz range, which overlaps significantly with the frequency spectrum of most music. This presents challenges for accurate respiratory rate estimation, as noted in previous studies [13]. Addressing this issue may require hardware-level enhancements and signal preprocessing. For example, subtracting the music signal from the microphone input could help isolate respiratory sounds, enabling more reliable estimation. Inspired by reference [32], we can also explore the use of reconstructed ECG signals for respiratory rate estimation based on the principle of respiratory sinus arrhythmia (RSA). RSA reflects the modulation of heart rate variability by respiratory activity. Therefore, by analyzing the trend of heart rate variations in the reconstructed ECG signal, it is possible to achieve effective estimation of the respiratory rate.

Finally, in this work, we focus primarily on physiological signal monitoring during aerobic exercises, and do not account for performance under speaking conditions. In future work, we will incorporate more activity categories and conduct experiments in more complex scenarios, aiming to further improve the system's robustness. This will ensure that the system performs effectively across diverse real-world conditions, thus enhancing its reliability and adaptability across a broader range of activities.

VIII. Conclusion

This paper designs an earable physiological sensing system based on an in-ear microphone, VitalEar, which enables HR and RR estimation during various aerobic exercises. For HR estimation, VitalEareffectively separates the heartbeat period and harmonic coefficients, using them as additional features combined with deep learning techniques to achieve accurate HR estimation. For RR estimation, VitalEarconstructs spatial and temporal feature extraction modules, effectively addressing the diversity issues introduced by different users and exercise intensities, while overcoming background noise interference. And by capturing capture the long-term dependencies of breathing, VitalEarachieves robust RR estimation. Through extensive experiments and mobile deployment, we demonstrate the effectiveness of VitalEarin HR and RR estimation, providing new insights for the research and application of earable devices.

References

[1] American College of Sports Medicine, "Position statement on the recommended quantity and quality of exercise for developing and maintaining fitness in healthy adults," *Med. Sci. Sports Exer.*, vol. 10, pp. vii–x, 1978.

[2] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The importance of respiratory rate monitoring: From healthcare to sport and exercise," *Sensors*, vol. 20, no. 21, 2020, Art. no. 6396.

[3] F. O. Pires et al., "Cardiopulmonary, blood metabolite and rating of perceived exertion responses to constant exercises performed at different intensities until exhaustion," *Brit. J. Sports Med.*, vol. 45, no. 14, pp. 1119–1125, 2011.

[4] S. Marcora, "Counterpoint: Afferent feedback from fatigued locomotor muscles is not an important determinant of endurance exercise performance," *J. Appl. Physiol.*, vol. 108, no. 2, pp. 454–456, 2010.

[5] A. Nicolò, C. Massaroni, and L. Passfield, "Respiratory frequency during exercise: The neglected physiological measure," *Front. Physiol.*, vol. 8, 2017, Art. no. 922.

[6] A. Nicolò, M. Montini, M. Girardi, F. Felici, I. Bazzucchi, and M. Sacchetti, "Respiratory frequency as a marker of physical effort during high-intensity interval training in soccer players," *Int. J. Sports Physiol. Perform.*, vol. 15, no. 1, pp. 73–80, 2020.

[7] M. Girardi, A. Nicolò, I. Bazzucchi, F. Felici, and M. Sacchetti, "The effect of pedalling cadence on respiratory frequency: Passive vs. active exercise of different intensities," *Eur. J. Appl. Physiol.*, vol. 121, pp. 583–596, 2021.

[8] C. M. Albert, M. A. Mittleman, C. U. Chae, I.-M. Lee, C. H. Hennekens, and J. E. Manson, "Triggering of sudden death from cardiac causes by vigorous exertion," *New England J. Med.*, vol. 343, no. 19, pp. 1355–1361, 2000.

[9] X. Jouven, J.-P. Empana, P. J. Schwartz, M. Desnos, D. Courbon, and P. Ducimetière, "Heart-rate profile during exercise as a predictor of sudden death," *New England J. Med.*, vol. 352, no. 19, pp. 1951–1958, 2005.

[10] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *Int. J. Biosensors Bioelectron.*, vol. 4, no. 4, pp. 195–202, 2018.

[11] ZephyrTechnology, "Bioharness 3.0 user manualbioharness 3.0 user manual," Aug., 2024. [Online]. Available: https://www.habdirect.com/wp-content/uploads/2018/08/bioharness3-user-manual.pdf

[12] A. Martin and J. Voix, "In-ear-audio-wearable: Measurement of heart and breathing rates for health and safety monitoring," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1256–1263, Jun. 2018.

[13] T. Ahmed, M. M. Rahman, E. Nemati, M. Y. Ahmed, J. Kuang, and A. J. Gao, "Remote breathing rate tracking in stationary position using the motion and acoustic sensors of earables," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–22.

[14] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, and C. Mascolo, "hEARt: Motion-resilient heart rate monitoring with in-ear microphones," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2023, pp. 200–209.

[15] C. Hu, T. Kandappu, Y. Liu, C. Mascolo, and D. Ma, "BreathPro: Monitoring breathing mode during running with earables," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 1–25, 2024.

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

[17] V. I. Korenbaum et al., "Human forced expiratory noise. origin, apparatus and possible diagnostic applications," *J. Acoust. Soc. Amer.*, vol. 148, no. 6, pp. 3385–3391, 2020.

[18] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London Ser. A: Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[19] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, Aug., 2013.

[20] I. Daubechies and C. Heil, "Ten lectures on wavelets," *Comput. Phys.*, vol. 6, no. 6, 1992, Art. no. 697.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf., Med. Image Comput., Comput.-Assist. Interv.*, Munich, Germany, Oct., 2015, pp. 234–241.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[23] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[24] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.

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

[26] Y. Jin, M. M. Rahman, T. Ahmed, J. Kuang, and A. J. Gao, "RRDetection: Respiration rate estimation using earbuds during physical activities," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2023, pp. 1–5.

[27] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, Y. Liu, and C. Mascolo, "An evaluation of heart rate monitoring with in-ear microphones under motion," *Pervasive Mobile Comput.*, vol. 100, 2024, Art. no. 101913.

[28] X. Sun et al., "Earmonitor: In-ear motion-resilient acoustic sensing using commodity earphones," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–22, 2023.

[29] X. Fan, D. Pearl, R. Howard, L. Shangguan, and T. Thormundsson, "APG: Audioplethysmography for cardiac monitoring in hearables," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw.*, 2023, pp. 1–15.

[30] Q. Zhang, X. Zeng, W. Hu, and D. Zhou, "A machine learning-empowered system for long-term motion-tolerant wearable monitoring of blood pressure and heart rate with ear-ECG/PPG," *IEEE Access*, vol. 5, pp. 10547–10561, 2017.

[31] T. Ahmed, M. M. Rahman, E. Nemati, M. Y. Ahmed, J. Kuang, and A. J. Gao, "Remote breathing rate tracking in stationary position using the motion and acoustic sensors of earables," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–22.

[32] Y. Liu, K.-J. Butkow, J. Stuchbury-Wass, A. Pullin, D. Ma, and C. Mascolo, "RespEar: Earable-based robust respiratory rate monitoring," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2024, pp. 67–77.

[33] D. Liaqat et al., "WearBreathing: Real world respiratory rate monitoring using smartwatches," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–22, 2019.

[34] P.-Y. Hsu, P.-H. Hsu, T.-H. Lee, and H.-L. Liu, "Heart rate and respiratory rate monitoring using seismocardiography," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 6876–6879.

[35] Y. Cao, F. Li, H. Chen, X. Liu, L. Zhang, and Y. Wang, "Guard your heart silently: Continuous electrocardiogram waveform monitoring with wrist-worn motion sensor," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–29, 2022.

[36] X. Guo et al., "Exploring biomagnetism for inclusive vital sign monitoring: Modeling and implementation," in *Proc. 30th Annu. Int. Conf. Mobile Comput. Netw.*, 2024, pp. 93–107.

[37] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *NPJ Digit. Med.*, vol. 3, no. 1, 2020, Art. no. 18.

[38] Y. Yang, J. Cao, and X. Liu, "ER-Rhythm: Coupling exercise and respiration rhythm using lightweight COTS RFID," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 4, pp. 1–24, 2019.

[39] R. Adhikary et al., "Spiromask: Measuring lung function using consumer-grade masks," *ACM Trans. Comput. Healthcare*, vol. 4, no. 1, pp. 1–34, 2023.

[40] C. Romano et al., "Respiratory rate estimation during walking and running using breathing sounds recorded with a microphone," *Biosensors*, vol. 13, no. 6, 2023, Art. no. 637.

[41] W. Xie, Q. Hu, J. Zhang, and Q. Zhang, "Earspiro: Earphone-based spirometry for lung function assessment," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–27, 2023.

[42] Y. Zou, J. Weng, H. Lei, D. Wang, V. C. M. Leung, and K. Wu, "EarPrint: Earphone-based implicit user authentication with behavioral and physiological acoustics," *IEEE Internet Things J.*, vol. 11, no. 19, pp. 31128–31143, Oct., 2024.

[43] Y. Cao, C. Cai, F. Li, Z. Chen, and J. Luo, "HeartPrint: Passive heart sounds authentication exploiting in-ear microphones," in *Proc. IEEE Conf. Comput. Commun.*, 2023, pp. 1–10.

[44] P. Zhu, Y. Zou, W. Li, and K. Wu, "CHAR: Composite head-body activities recognition with a single earable device," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2023, pp. 212–221.

[45] Y. Cao, C. Cai, A. Yu, F. Li, and J. Luo, "EarACE: Empowering versatile acoustic sensing via earable active noise cancellation platform," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 2, pp. 1–23, 2023.

[46] D. Ma, A. Ferlini, and C. Mascolo, "OESense: Employing occlusion effect for in-ear human sensing," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, 2021, pp. 175–187.

Yuzheng Zhu received the master's degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in June 2025. He conducted this work when he was working toward the master's degree with Shenzhen University. Since 2025, he is has been an algorithm engineer with DJI. His research interests include earable computing and intelligent sensing.

Zhangxin Liang received the bachelor's degree in 2024 from the College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China, where he is currently working toward to the master's degree with the College of Computer Science and Software Engineering. His research interests include intelligent sensing and earable computing.

Jie Zheng is currently working toward the bachelor's degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests include affective computing and intelligent sensing.

Yongpan Zou (Member, IEEE) received the PhD degree from the Department of Computer Science and Engineering (CSE), Hong Kong University of Science and Technology, Hong Kong, in 2017. He is currently an associate professor with the College of Computer Science and Software Engineering, Shenzhen University. His research interests include ubiquitous sensing, mobile computing, and human-computer interaction.

Victor C. M. Leung (Life Fellow, IEEE) received the PhD degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, in 1981. He is currently the dean with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China. His research interests include wireless networks and mobile systems. He was the recipient of the numerous accolades, such as the APEBC Gold Medal, NSERC Postgraduate Scholarships, and IEEE Vancouver Section Centennial Award. He has more then 60 000 citations.

Kaishun Wu (Fellow, IEEE) received the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2011. He is currently a professor in information hub with the Hong Kong University of Science and Technology (Guangzhou). His research interests include wireless communications and mobile computing. He was the recipient of the several Best Paper awards of international conferences, such as IEEE Globecom 2012 and IEEE MASS 2014.