EarSleep: In-ear Acoustic-based Physical and Physiological Activity Recognition for Sleep Stage Detection

FEIYU HAN, University of Science and Technology of China, China
PANLONG YANG, Nanjing University of Information Science and Technology, China
YUANHAO FENG, The Hong Kong Polytechnic University, China
WEIWEI JIANG, Nanjing University of Information Science and Technology, China
YOUWEI ZHANG, The University of Electro-Communications, Japan
XIANG-YANG LI, University of Science and Technology of China, China

Since sleep plays an important role in people's daily lives, sleep monitoring has attracted the attention of many researchers. Physical and physiological activities occurring in sleep exhibit unique patterns in different sleep stages. It indicates that recognizing a wide range of sleep activities (events) can provide more fine-grained information for sleep stage detection. However, most of the prior works are designed to capture limited sleep events and coarse-grained information, which cannot meet the needs of fine-grained sleep monitoring. In our work, we leverage ubiquitous in-ear microphones on sleep earbuds to design a sleep monitoring system, named EarSleep[1], which interprets in-ear body sounds induced by various representative sleep events into sleep stages. Based on differences among physical occurrence mechanisms of sleep activities, EarSleep extracts unique acoustic response patterns from in-ear body sounds to recognize a wide range of sleep events, including body movements, sound activities, heartbeat, and respiration. With the help of sleep medicine knowledge, interpretable acoustic features are derived from these representative sleep activities. EarSleep leverages a carefully designed deep learning model to establish the complex correlation between acoustic features and sleep stages. We conduct extensive experiments with 48 nights of 18 participants over three months to validate the performance of our system. The experimental results show that our system can accurately detect a rich set of sleep activities. Furthermore, in terms of sleep stage detection, EarSleep outperforms state-of-the-art solutions by 7.12% and 9.32% in average precision and average recall, respectively.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Sleep Monitoring, Sleep Stage Detection, In-ear Body Sound, Earable Sensing

ACM Reference Format:
Feiyu Han, Panlong Yang, Yuanhao Feng, Weiwei Jiang, Youwei Zhang, and Xiang-Yang Li. 2024. EarSleep: In-ear Acoustic-based Physical and Physiological Activity Recognition for Sleep Stage Detection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 43 (June 2024), 31 pages. https://doi.org/10.1145/3659595

[1]The demo video is available at https://www.youtube.com/watch?v=23MpIv_BaVc

Authors' Contact Information: Feiyu Han, University of Science and Technology of China, Hefei, China, fyhan@mail.ustc.edu.cn; Panlong Yang, Nanjing University of Information Science and Technology, Nanjing, China, plyang@nuist.edu.cn; Yuanhao Feng, The Hong Kong Polytechnic University, Hong Kong, China, yuanhfeng@polyu.edu.hk; Weiwei Jiang, Nanjing University of Information Science and Technology, Nanjing, China, weiweijiangcn@gmail.com; Youwei Zhang, The University of Electro-Communications, Tokyo, Japan, zhanguv@uec.ac.jp; Xiang-Yang Li, University of Science and Technology of China, Hefei, China, xiangyangli@ustc.edu.cn.

ACM 2474-9567/2024/5-ART43
https://doi.org/10.1145/3659595

1 INTRODUCTION

Sleep health is considered an under-recognized global public health issue and has become one of the risk factors that seriously threaten public health [18, 21]. Long-term sleep monitoring plays a crucial role in understanding and assessing various aspects of sleep health, which attracts more attention. The assessment of sleep health requires careful measurement of sleep stages. Thus, it is crucial to detect and track sleep stages accurately for sleep monitoring. Prior sleep-medicine studies [1, 31, 35, 56] have reported that human activities occurring in sleep are related to sleep stage transitions. These sleep activities can be divided into two categories: observable physical activities (*e.g.*, body movements and sound activities) and hardly-observable physiological activities (*e.g.*, breathing, heartbeat, eye movements, and brain activities). We refer to these human activities that occur during sleep as sleep events. The key insight of sleep stage detection is that people usually exhibit unique sleep-activity patterns in different sleep stages [6, 17, 35].

Traditionally, polysomnography (PSG) is considered the "gold standard" of sleep study [4], which measures EEG, EMG, EOG, ECG, and others for comprehensive assessment of sleep. However, the complex operations (multiple wire/electrode attachments to the body) and high costs (ranging from 1000$ to 7000$ [66]) limit its use to clinical trials only. Recently, dedicated biosensors *e.g.*, photoplethysmography (PPG) sensor and electrocardiogram (ECG) sensor, are designed to collect variations in physiological signals for sleep monitoring, but they are not widely available on most devices due to high cost and integration complexity. Therefore, the academic community has shown substantial interest in portable home sleep monitoring that utilizes available and ubiquitous sensors on commercial devices. To capture a wide range of sleep events containing sleep-related information, prior works [13, 19] utilize multiple sensors on mobile phones to recognize most physical activities (*e.g.*, sound activities and body movements). Furthermore, Chang *et al.* [12] utilize multiple sensors on wrist-worn devices to capture both physical activities and important physiological activities for sleep monitoring. Although these significant works have greatly advanced ubiquitous sleep monitoring, they still have limitations in fine-grained sleep monitoring. Firstly, most of them only detect limited sleep events that are insufficient to provide fine-grained information for sleep monitoring. For instance, Sleep Hunter [19] only captures most physical activities occurring in sleep while ignoring important physiological activities and SleepGuard [12] cannot capture variations in heartbeat. Second, they don't fully exploit the intra- and inter-characteristics of sleep events but only use the waveform-domain statistical characteristics for sleep monitoring, which is still unable to achieve fine-grained sleep stage detection. Lastly, most of them require multiple sensors embedded in the monitoring devices to recognize different sleep events, which leads to additional hardware costs or complex engineering integration.

To solve the above limitations, we design EarSleep, an in-ear acoustic-based sleep monitoring system that achieves fine-grained sleep stage detection via a wide range of representative sleep events. EarSleep is built on a pair of commercial sleep earbuds with in-ear microphones. The key observation behind the system design is that body vibrations induced by human activities propagate through bone conduction to the ear canal and can be captured by the in-ear microphone. As shown in Fig. 1, EarSleep only leverages in-ear microphones on the earbuds to recognize representative physical activities (including body movements and sound activities) and physiological activities (including heartbeat and breathing) occurring in sleep. Based on the correlation between these sleep events and sleep stages, EarSleep extracts effective physical and physiological features for accurate sleep stage detection. Different from prior multiple sensor-based approaches, EarSleep only takes advantage of a single in-ear acoustic modality to recognize a wide range of representative sleep events that contain sleep-related information, and translate these sleep events into sleep stages based on the interpretable correlation.

Technical challenges. To design such a promising earbud-based sleep monitoring system, we need to address several core technical challenges as follows:

i) How to accurately recognize various physical activities during sleep without assistance from other modalities? Sleep is a continuous and long-time process, where various types of sleep events occur. As introduced in Sec. 3.1,

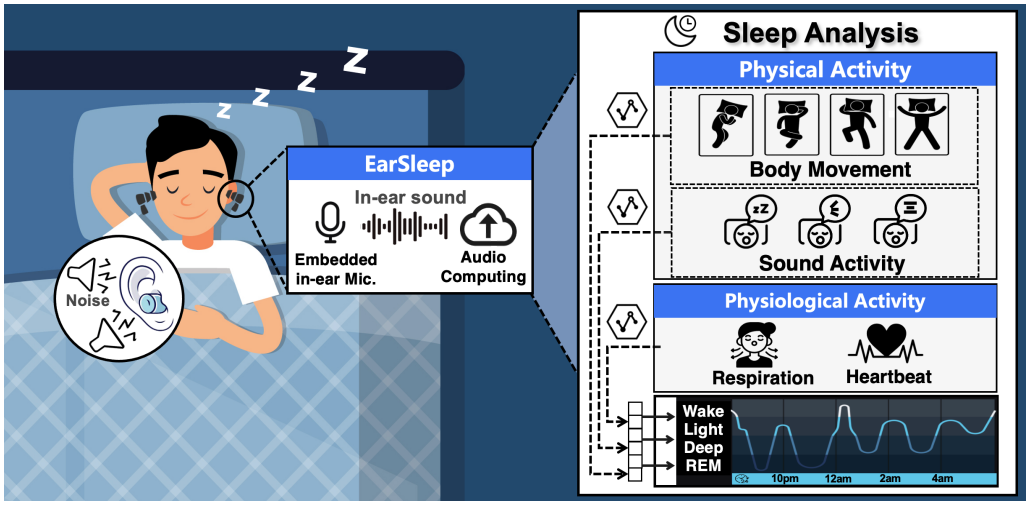


Fig. 1. The goal of EarSleep and application scenarios.

sleep events have various patterns such as intensity, duration, and periodicity. Previous solutions [12, 19] use multiple modalities/sensors to identify different types of sleep events. However, there is only one sensing modality (*i.e.*, in-ear sounds) available to us, making it challenging to identify these events with high diversity. To address this challenge, we design a series of sleep event detection and segmentation methods according to various sleep events' unique in-ear acoustic responses. Then, *combing physical occurrence mechanisms of sleep events with behavior statistical analysis*, discriminate acoustic representations are extracted to recognize four types of body movements (*i.e.*, body rollover, body trembling, turning head, and limb movement) and three types of sound activities (*i.e.*, snoring, coughing, and somniloquy) that are related to sleep health and movement disorders [2, 36].

ii) How to obtain accurate physiological activity estimation in the presence of motion artifacts? Compared with physical activities, heartbeat and breathing are more tiny physiological activities in sleep. As introduced in Sec. 3.3, heartbeat and breathing in-ear sounds are heavily disrupted by motion artifacts. To address this challenge, we propose a VMD-based decomposition method to extract the pure heartbeat and breathing waveforms from noisy in-ear sounds. The waveform with maximum HNR (heartbeat-to-noise ratio) and PNR (periodicity-to-noise ratio), and the waveform with maximum BNR (breathing-to-noise ratio) and PNR are selected to perform heartbeat and breathing information (*e.g.*, rate and interval) estimation, respectively.

iii) How to associate various sleep events with sleep stages via representative and interpretable acoustic features? After detecting all sleep events, establishing the correlation between sleep events and sleep stages is challenging since there are no existing well-developed solutions utilizing the in-ear acoustic modality for sleep monitoring. Prior works [5, 38, 56] in sleep medicine have studied the variation patterns of physiologic parameters in each sleep stage, which can inspire us to integrate sleep medicine knowledge with in-ear acoustic sleep sensing. Specifically, we extract fine-grained acoustic features from in-ear sounds to represent the variation patterns of physical and physiological parameters. In addition, the guidance of knowledge of sleep medicine improves the interpretability of these sleep-related acoustic features. Then these features are input to a carefully-designed BiLSTM-based model with an attention mechanism for sleep stage inference.

In summary, this work makes the following contributions:

- As far as we know, we are the first to present an in-ear acoustic-based sleep monitoring system that only utilizes an in-ear microphone to capture a wide range of sleep event information and achieve fine-grained

sleep stages detection. Our work can help people understand the impact of sleep events on sleep health and provide a fine-grained assessment of sleep in a ubiquitous way.
- We present a comprehensive unique challenge analysis for in-ear acoustic-based long-term sleep monitoring in practice (see Sec. 3). To address these challenges, we develop a dual-stream processing framework (see Sec. 4) to simultaneously recognize physical and physiological activities occurring in sleep (see Sec. 5.2 and Sec. 5.3). Based on the correlation between sleep events and sleep stages, EarSleep derives acoustic representations from a wide range of sleep events for sleep stage detection (see Sec. 5.4).
- Through extensive real-world experiments (see Sec. 6), we have demonstrated that our system can achieve accurate four-class body movement recognition, three-class sound activity recognition, and physiological activity estimation. Compared with existing mobile/wearable solutions, EarSleep can also yield an outstanding performance of sleep stage detection.

2 MOTIVATION AND BACKGROUNDS

2.1 Earbud for Sleep: Opportunity

With the expansion of the sleep economy market, the sleep earbuds market is valued at approximately $15 million in 2020 and will continue to grow in the coming years [26]. The emerging trend of earable sleep technology benefits from the following reasons. Typically, different from wrist-worn and body-worn devices, sleep earbuds are worn in ear canals which are the ideal positions for measuring physiological parameters that are associated with sleep stages [55]. In addition, there are an increasing number of technological advancements in noise cancellation, ergonomic shape design, soft silicone material, and others that aim to improve comfortability during overnight wearing. For instance, the ergonomic in-ear design ensures that earbuds can block out external noises into ear canals, and the attachment designs (such as hook around the ear and concha filling) ensure that earbuds will not fall off easily during sleep. However, existing commercial sleep earbuds (*e.g.*, Bose Sleepbuds II and Amazfit ZenBuds) mainly rely on specially-designed biological sensors (*e.g.*, ECG and PPG) for sleep monitoring, making the cost expensive (above 100 $ [52]). Thus, we aim to explore the feasibility of utilizing available and ubiquitous sensors on earbuds for individual sleep monitoring.

2.2 Relationship between Sleep Stages and Sleep Events

People's sleep consists of several cycles (about 4-5) and each sleep cycle can be further divided into three stages, *i.e.*, rapid eye movement (REM) stage, light sleep, and deep sleep [47] that follow a predictable pattern. It indicates that there are dependencies or connections between sleep stages before and after the current moment [47]. The inherent context connections among sleep stages can help us detect the sleep stages. Distinguishable physical activities and physiological activities are also exhibited in different sleep stages [6, 17, 35]. In the light sleep stage, the body starts to relax gradually, and most of the physiological activities (*e.g.*, heartbeat, breathing, and brain activities) start to slow down. In the deep sleep stage, physiological activities will be further decreased compared with the light sleep stage. Light body movements (*e.g.*, body trembling, arm and leg jerking, and head movement) and snoring mostly occur in this stage. In the REM sleep stage, some physiological activities like brain activities and eye movements, become more frequent, which may lead to variations in heartbeat rate and breathing rate. It indicates that the variations in sleep events can be considered as indicators of sleep stages [1, 56, 59].

2.3 In-ear Body Sounds Acquisition

As the person wears earbuds, the ear canal is occluded by the earbud, which contributes to a blocked space between the eardrum and earbud. The part of the body where the activity is generated acts as a source of vibration. As a common physiological phenomenon, body conduction is the way sound travels along the bones of the body [67]. Therefore, the vibration signals induced by activities are transmitted to the inner ear via bone conduction. Due to the occluded space between the earbud and the eardrum, the body-conducted vibrations will be trapped and

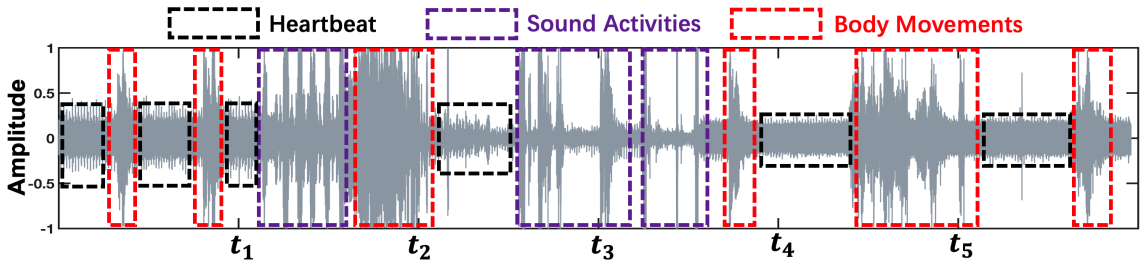


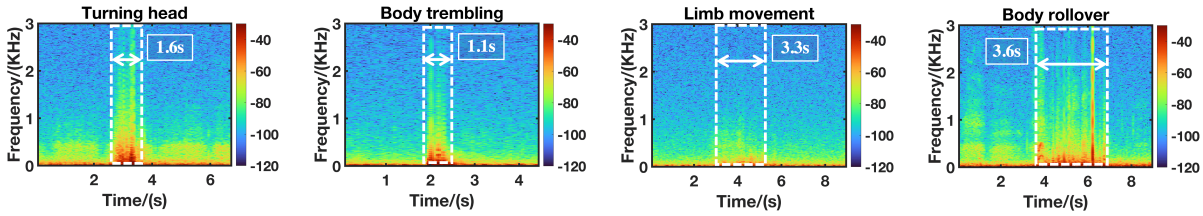
Fig. 2. Illustration of sleep events over a period of time.

reflected. The inward-facing microphone on the earbud can capture these echo-like vibrations/sounds which are called in-ear sounds. In addition, the low-frequency components (below 1000 Hz) of in-ear sounds can be significantly enhanced benefiting from the occlusion effect [10, 37], which ensures the reliability of the in-ear acoustic response of various sleep activities.

3 IN-EAR SOUNDS OF VARIOUS SLEEP EVENTS.

3.1 Diversity Analysis in Time domain.

Sleep events consist of various physical activities and physiological activities that have different patterns such as periodicity, duration, frequency, and intensity. As for essential physiological activities, heartbeat and breathing are long-time and periodic sleep events. Most body movements and sound activities are short-time and non-periodic but may occur many times during sleep. We segment a part of sleep events from a subject's sleep throughout the night which are shown in Fig. 2. We can observe that the amplitude of heartbeat-induced sound ranges from $[-0.4, 0.4]$, which is smaller than body-movement-induced sound and sound-activity-induced sound. However, sound activities and body movements are not distinguishable along the time domain, as their amplitudes both range from $[-1, 1]$. Thus, accurate segmentation and classification of sleep events with high diversity throughout the night are critical for fine-grained sleep monitoring.



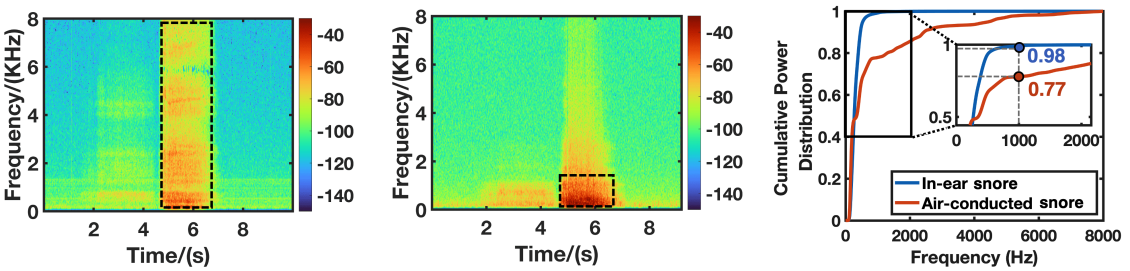
(a) Turning head. (b) Body trembling. (c) Limb movement. (d) Body rollover.

Fig. 3. The spectrograms of four types of body movements.

3.2 Physical Activity-induced Sound in Time-frequency Domain

3.2.1 ***Body Movement-induced In-ear Sound****. Body movements are completed by different parts of the body, and there is overlap between different movements (e.g., body rollover contains head and limb movements), resulting in only subtle differences in acoustic responses of body movements.* The limb movement, head movement, body trembling, and body rollover are the most frequent body movements that occur during sleep. These movements are associated with not only sleep disorders but also movement disorders such as head jerks [36] and restless legs syndrome (RLS) [2]. According to the duration, we divide body movements into micro movements (*i.e.*, head

movement and body trembling) and macro movements (*i.e.*, limb movement and body rollover). However, the difference among durations of sleep events is not sufficient for event classification. Next, we need to analyze the differences in the time-frequency domain, as shown in Fig. 3. Body rollover has the longest duration and highest power spectral density (PSD). However, as for a macro body movement, the PSD of in-ear sound induced by limb movement is less than micro movements. We combine the physical occurrence mechanisms of these body movements to explain this phenomenon. Different body movements involve different parts of the body and have unique physical occurrence mechanisms. The limb movement involves hands and legs working together. The body rollover and body trembling involve the whole body (including hands, legs, and head). As we all know, the human body is a dispersive medium that makes movement-induced sounds attenuated. Compared with the head, limbs are farther away from the ears, which causes the limb movement-induced sound to travel a longer path to reach the in-ear microphone, leading to severe signal attenuation. The body trembling and body rollover both involve the head movement, thus these body movements have a larger power spectral density.



(a) **Air-conducted snore sound.** (b) **In-ear snore sound.** (c) **Cumulative power distribution.**

Fig. 4. Comparison of the snoring sound of two channels.

3.2.2 ***Sound Activity-induced In-ear Sound****. The occlusion effect enhances the low-frequency components of in-ear sounds, which changes the structure of spectrogram.* Snoring, coughing, and somniloquy are three common types of sound activities. Prior works [12, 19, 53, 69] utilize the microphone on the mobile device or wearable device to capture air-conducted sounds for sound activity detection during sleep. We take snoring as an example to study the differences between air-conducted sounds and body-conducted sounds. As shown in Fig. 4(a) and Fig 4(b), most of the energy of in-ear snore sound is concentrated in the low-frequency part compared with the air-conducted snore sound. For example, about 98% of power is distributed below 1 KHz for the in-ear snore sound, while only about 77% of power is distributed below 1 KHz for the air-conducted snore sound, as shown in Fig. 4(c). That is because the occlusion effect significantly enhances the low-frequency (below 1 KHz) components of body sounds [10], causing the spectrogram of in-ear snoring sound to be different from that of air-conducted snoring sound. The formant pattern is a widely used acoustic feature in speech science and phonetics [71]. We use the t-SNE method [68] to visualize the formant patterns that are extracted from three types of sound activities. As shown in Fig. 5, we can observe that formant-related features are not sufficient to distinguish different events. Hence, we should also select more discriminable features that are adapted to in-ear acoustic modality, to represent each type of sound activity during long-time sleep.

3.3 Artifacts Analysis of Physiological Activity-induced Sound

Body movement-induced and sound activity-induced in-ear sounds severely interfere with the physiological activity-induced in-ear sounds. Compared with body movements and sound activities, breathing and heartbeat are subtle physiological activities. As shown in Fig. 2, physiological activity-induced in-ear sound has a smaller amplitude

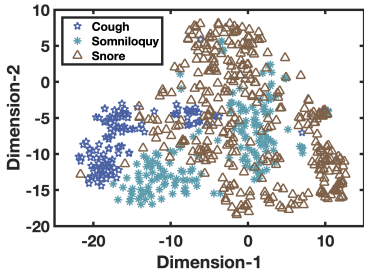


Fig. 5. t-SNE visualization of three types of sound activities.

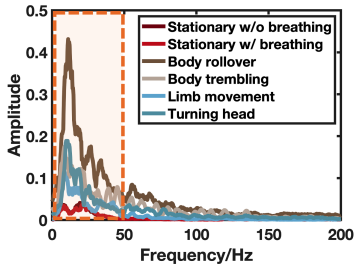


Fig. 6. Impact of body movement on the physiological activity-induced sound.

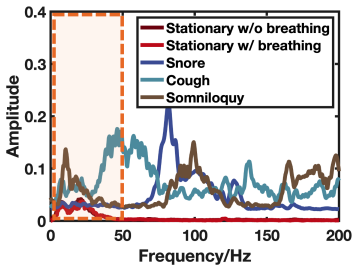


Fig. 7. Impact of sound activity on the physiological activity-induced sound.

range compared with physical activity-induced sounds. We next select several typical sleep events to give the analysis of artifacts on in-ear sounds of physiological activities.

Artifacts on heartbeat sounds. When the user is stationary without breathing, there are only heartbeat-induced in-ear sounds. As shown in Fig. 6 and Fig. 7, we can observe that the frequency range of heartbeat sound is mainly distributed between $[0, 50]$ Hz (*i.e.*, the orange box with dot line), which overlaps with the frequency ranges of body movement-induced sound and sound activity-induced sound. It indicates that the low-frequency components of body movements and sound activities generate artifacts on the in-ear heartbeat sound.

Artifacts on breathing sounds. When the user is stationary with breathing, there are mixed body sounds induced by heartbeat and breathing. From Fig. 6 and Fig. 7, we can both observe that the mixed spectral energy of heartbeat and breathing is similar to the spectral energy of heartbeat. It indicates that the acoustic response of the heartbeat is stronger than that of breathing. The breathing sounds are essentially weak vibrations produced by the collision and friction of the inhaled or exhaled airflow with the airway walls [57], which are mainly distributed below 3 KHz [24]. Therefore, compared with sounds induced by other sleep events, breathing-induced in-ear sounds have lower SNR which are easily overwhelmed by noises.

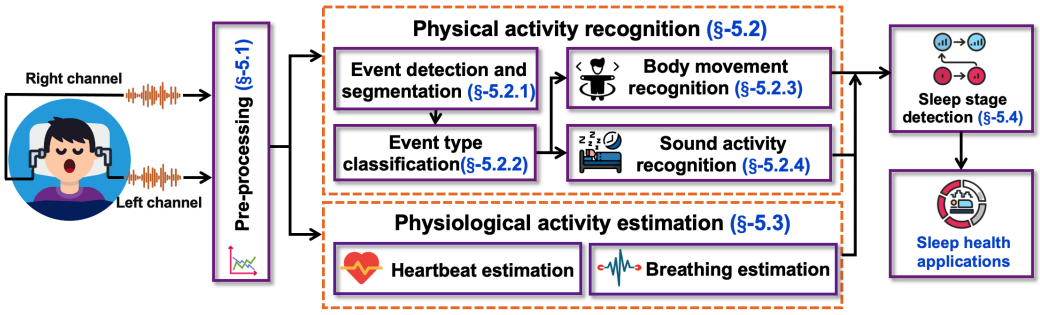


Fig. 8. Design overview of EarSleep. A dual data-stream framework is designed to simultaneously recognize physical activities related to sleep and estimate the sleep information of physiological activities.

4 SYSTEM OVERVIEW

The overview of EarSleep is illustrated in Fig. 8. EarSleep utilizes in-ear microphones on sleep earbuds to automatically capture two-channel (*i.e.*, left and right) body sounds induced by various sleep events. Then, after pre-processing (Sec. 5.1), the acoustic signals are fed into a dual data-stream framework for parallel processing.

Physical activity recognition (Sec. 5.2). In the one stream, EarSleep detects and separates each physical activity as an acoustic event by adaptive segmentation methods (Sec. 5.2.1). EarSleep identifies whether the event type is a body movement or a sound activity (Sec. 5.2.2). According to the result of event type classification, the sleep event is further input to the corresponding fine-grained recognition module, *i.e.*, body movement recognition module (Sec. 5.2.3) or sound activity recognition module (Sec. 5.2.4).

Physiological activity recognition (Sec. 5.3). In another parallel stream, EarSleep extracts the heartbeat estimation and breathing estimation from the in-ear body sound based on characteristics in frequency components.

Sleep stage detection (Sec. 5.4). Finally, EarSleep extracts sleep-related features from above acoustic sleep activities (Sec. 5.4.1). A deep-learning-based model is constructed to represent the relationship between various sleep events and sleep stages and fuses features of various sleep events to identify sleep stages (Sec. 5.4.2).

5 SYSTEM DESIGN

5.1 Pre-processing

Considering that sleep is a continuous process, it makes sense for us to analyze sleep events from in-ear body sounds for a period of time. Therefore, in-ear body sounds will be stored in the buffer with a length of 60 seconds. When the buffer is full, EarSleep will take out all acoustic signals in the buffer for pre-processing. We first normalize in-ear acoustic signals with the min-max normalization method to eliminate individual differences. Then, normalized acoustic signals are subtracted by the average of normalized acoustic signals to remove the DC offset introduced by the microphone hardware. In practice, we find that there are some outliers (appearing as spikes) caused by circuits of microphones in acoustic signals. Hence, we continue to use the Hampel filter to remove these outliers by replacing outliers with the local median.

5.2 Physical Activity Recognition

Physical activities occurring in sleep mainly consist of various body movements and sound activities. In particular, body movements include head turning, body rollover, limb movement, and body trembling, and sound activities include snoring, coughing, and talking in the dream (somniloquy). As described in Sec. 3.2, physical activity-induced sounds are distributed below 4 KHz. Thus, we first downsample pre-processed audio signals to 8 KHz.

5.2.1 ***Event Detection and Segmentation.*** We refer to a physical activity as an event. As introduced in Sec. 3.1, there are various sleep events with high diversity during the night, which makes conventional threshold-based methods unsuitable for our work. Thus, our primary goal is to construct an adaptive and computing-efficient method for event detection and segmentation.

(i) Remove Impact of Heartbeat. As analyzed in Sec. 3.3, in-ear heartbeat sound interferes with non-physiological event detection and segmentation. Thus, we first use the high-pass filter with a cut-off frequency of 100 Hz to eliminate the impact of heartbeat sound. As the spectral power of breathing-induced sound is much smaller than that of non-physiological activities, we do not need to consider the impact of breathing noise on event detection and segmentation.

(ii) Envelope Extraction with Logarithm and Linear Function Scaling. We employ a sliding hamming window with a length of 0.2 seconds and a step with a length of 0.1 seconds to split acoustic signals into N frames. As for the k-th frame x_k with l samples, we calculate the short-time energy which can be expressed as $STE(k) = \sum_{t=1}^{l} |x_k(t)|^2$. Then we extract the envelope of STE (denoted by Env_{rms}) using the root-mean-square function. Furthermore, to widen the differences between events and non-events, we rescale the RMS envelope

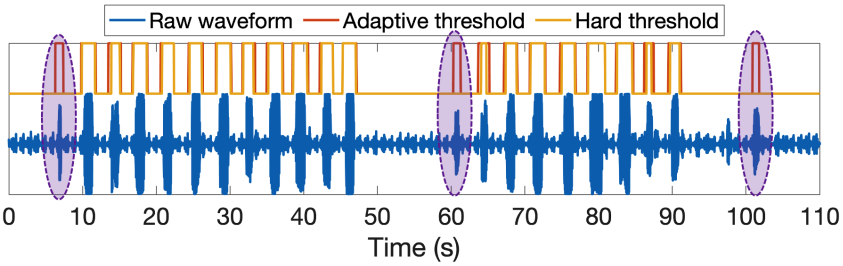


Fig. 9. Illustration of event segmentation using adaptive threshold and hard threshold.

function, which is given by

$$ReEnv_{rms}(k) = \begin{cases} \log_{T_{scale}} Env_{rms}(k) & \text{if} \quad Env_{rms}(k) < T_{scale} \\ \alpha(Env_{rms}(k) - T_{scale}) + 1 & \text{if} \quad Env_{rms}(k) \geq T_{scale} \end{cases} \tag{1}$$

where α and T_{scale} are the scaling factor and scaling threshold, respectively. T_{scale} determines that whether $Env_{rms}(k)$ is scaled or not. T_{scale} is related to the difference between events and non-events. Generally, the user rests on the bed with fewer body movements and sound activities within 1-3 minutes before falling asleep. Thus, we use the 3σ rule to initialize the scaling threshold T_{scale} with frames in the first three minutes, given by:

$$T_{scale} = mean(Env_{rms}(k)) + 3 * std(Env_{rms}(k)), \quad (k = 1, ..., 3 * N) \tag{2}$$

If $Env_{rms}(k) < T_{scale}$, the k-th frame is scaled towards negative infinity by the logarithm function. If $Env_{rms}(k) \geq T_{scale}$, the k-th frame is linearly amplified by positive scaling factor α which is set to 3 in our work.

(iii) Adaptive Event Segmentation. Then, we design an update mechanism to achieve adaptive segmentation. The key idea behind our update mechanism the current observation stage is related to previous observation stages, as sleep is a continuous process. Specifically, we define $\mathbf{S_{nE}}$ as the set of non-event frames and $\mathbf{S_E}$ as the set of event frames. Firstly, all frames of the first three minutes are added into $\mathbf{S_{nE}}$. Then, we set the initialization threshold T_{event} as $1.5 * T_{scale}$. Then we compare each sample of $ReEnv_{rms}$ with T_{event} to determine whether is an event or not. During the next processing, at time k, if $ReEnv_{rms}(k) < T_{event}$, the k-th frame is identified as the non-event, and we add $ReEnv_{rms}(k)$ into $\mathbf{S_{nE}}$. The threshold T_{event} should be updated based on the current observation and previous observations, which is given by

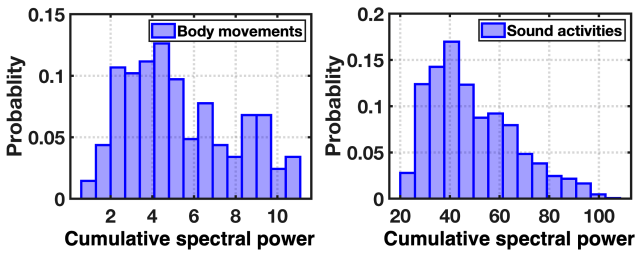
$$T_{event} = T_{event} + std(\mathbf{S_{nE}^k}) \quad \text{if } mean(\mathbf{S_{nE}^k}) + 2 * std(\mathbf{S_{nE}^k}) < ReEnv_{rms}(k) \tag{3}$$

where $\mathbf{S_{nE}^k} \subseteq \mathbf{S_{nE}}$ and can be expressed as $\mathbf{S_{nE}^k} = \{\mathbf{S_{nE}}(k - \Delta t), \mathbf{S_{nE}}(k - \Delta t + 1), ..., \mathbf{S_{nE}}(k)\}$. The Δt is set to 500. Otherwise, if the k-th frame is identified as the event, we do not need to update the threshold T_{event}. Above all, the update mechanism of T_{event} can be expressed as follows:

$$T_{event} = \begin{cases} 1.5 * T_{scale} & \text{if} \quad \text{initialization.} \\ T_{event} + std(\mathbf{S_{nE}^k}) & \text{if} \quad Env_{rms}(k-1) < T_{event} \text{ and} \\ & \quad mean(\mathbf{S_{nE}^k}) + 2 * std(\mathbf{S_{nE}^k}) < ReEnv_{rms}(k). \\ \text{not updated} & \text{if} \quad \text{others.} \end{cases} \tag{4}$$

Fig. 9 illustrates a time sequence where snoring occurs. We can observe that the segmentation method with the adaptive threshold can accurately detect the locations of sleep events. Since there are various sleep events with high diversity during sleep, the hard threshold-based segmentation method misses some minor events that are shown in the purple area in Fig. 9.

(iv) Duration Check. If the k-th frame is identified as the event, we add this frame into $\mathbf{S_E}$. After segmentation, there may be some incorrectly separated frames. Fortunately, the durations of sleep events and non-events still be distinct. We define multiple consecutive frames in $\mathbf{S_E}$ as an event candidate E. Then we calculate the duration



(a) Body movements. (b) Sound activities.

Fig. 10. Cumulative spectral power of physical activities.

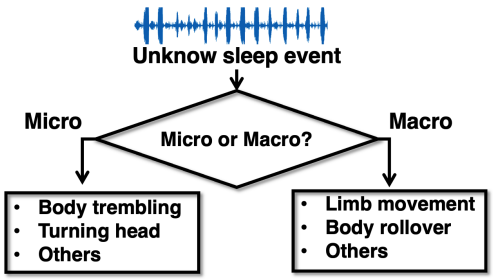


Fig. 11. Diagram of body movement recognition.

of each event candidate. If $E(i) < T_{min}$, we will drop it. T_{min} is the minimum duration of the event, which is set to 1 second. If the time interval between two consecutive frames is too close, *i.e.*, $|E(i) - E(i-1)| < T_{interval}$, we will merge $E(i-1)$ and $E(i)$ into one event. $T_{interval}$ is set to 0.5 seconds.

5.2.2 ***Event Type Classification: Is Body Movement or Sound Activity.*** After segmentation, EarSleep first performs the event type classification to determine whether each event is a body movement or a sound activity. As analyzed in Sec. 3.2.2 and Sec. 3.2.1, compared with the body movement-induced sound, the sound activity-induced sound has a strong acoustic response over a wider frequency range. Therefore, we use the power distribution of high-frequency components of in-ear sounds to discriminate body movements and sound activities. In particular, we find that power distribution patterns of body movements and sound activities are more discriminable above 200 Hz. Thus, we calculate the cumulative spectral power Eng_{cum} for frequency bins above 200 Hz of a sleep event. Fig. 10(a) and Fig. 10(b) show cumulative power distributions of body movements and sound activities. We can observe that power distributions of these two categories are significantly different. We adopt a cumulative power threshold of 15 to determine whether a sleep event is a sound activity or a body movement.

5.2.3 ***Body Movement Recognition.*** We mainly focus on four common movements that are related to sleep health, *i.e.*, turning head, body trembling, limb movement, and body rollover. We design an effective approach to analyze the acoustic representation of physical mechanisms behind the body movement occurrence for robust body movement recognition. The mechanism of body movement recognition is shown in Fig. 11.

(i) Unique Acoustic Response Analysis. We observed that both turning head and body rollover are completed by flipping some specific body parts (*e.g.*, head and trunk). We take an example of turning head in sleep. The structure of the head can be closely approximated using a simple geometry model which is presented in Fig. 12. When a subject turns his head towards to left, the head-to-pillow contact area is gradually close to the left ear, which results in the difference between the signal propagation path to the left ear d_l and the signal propagation path to the right ear d_r. As we all know, the human body can be considered as a dispersive medium that transmits acoustic signals of different frequencies with different attenuation [29, 51]. As the head frictions against the pillow during rotation, the vibration is generated at the head-to-pillow contact area and then travels along the skull to the ear. At the time t, we define that the vibration with the frequency f is $Vib(t, f)$. We model this signal propagation process as follows:

$$S_{in}(t, f) = H_{oe}(Vib(t, f) * e^{-(\alpha_r(f)+\alpha_s(f)+\alpha_a(f))*d}) \tag{5}$$

where $S_{in}(t, f)$ denotes body sounds captured by the in-ear microphone. Specifically, $H_{oe}, \alpha_r, \alpha_s, \alpha_a$, and d are the gain factor of the occlusion effect, reflection factor, scattering factor, absorption factor, and path length, respectively, which are related to individuals and can be considered as constants for simplicity. With the inspiration

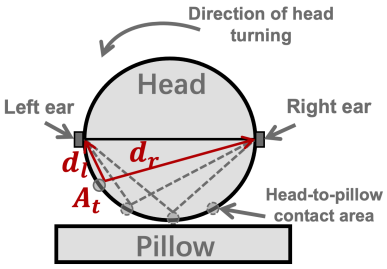


Fig. 12. Illustration of head movement.

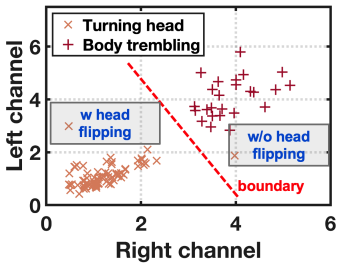


Fig. 13. 100∗zero-crossing rates of dual channels.

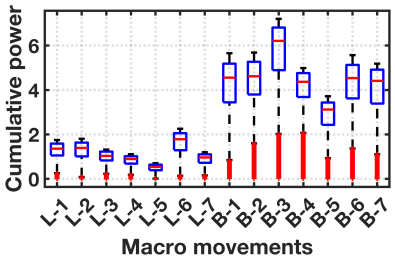


Fig. 14. The cumulative spectral power of macro movements.

of the ratio method [80], we can express the differences between left-channel and right-channel signals as follows:

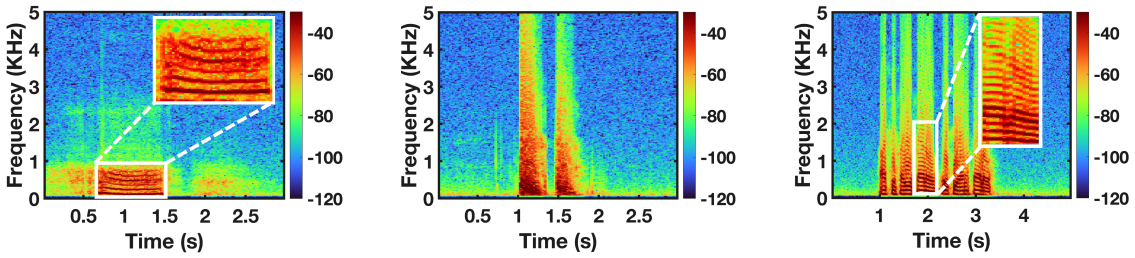
$$\frac{S_{inR}(t,f)}{S_{inL}(t,f)} = \frac{H_{oeR}(Vib(t,f) * e^{-(\alpha_r(f)+\alpha_s(f)+\alpha_a(f))*d_R})}{H_{oeL}(Vib(t,f) * e^{-(\alpha_r(f)+\alpha_s(f)+\alpha_a(f))*d_L})} \approx H * e^{(\alpha_r(f)+\alpha_s(f)+\alpha_a(f))*(d_L-d_R)} \tag{6}$$

Thus, we can observe that differences between left-channel and right-channel acoustic signals are mainly associated with d_l and d_r, *i.e.*, the distance from the vibration source to the left and right ears. The vibrations of different body movements travel to ears along different paths. Therefore, we can recognize different body movements via *acoustic attenuation of propagation paths* and *energy distribution induced by body parts*.

(ii) Is Micro or Macro Movement? Based on our long-time observation, we find that the head movement and body trembling are micro movements with shorter durations (ranging from 1 to 3 seconds) and the limb movement and body rollover are macro movements with longer durations (ranging from 5 to 10 seconds). Thus, we first identify whether the sleep event is a micro movement or a macro movement based on its duration.

(iii) Micro Movement Recognition: Turning head or Body trembling? Different propagation paths make different attenuation of each frequency component of the sound signal. Different from body trembling, turning head is a process of flipping the head with fewer movements of other parts of the body, which has unique attenuation profiles. Thus, we use the zero-crossing rate to measure the acoustic attenuation. We find that the zero-crossing rates of these two types of body movements are discriminable, as shown in Fig. 13. In addition, compared with turning head, body trembling involves more body parts (*e.g.*, head, limbs, and trunk), which causes the unique delay profiling between dual-channel acoustic signals. In our work, we split signals into frames. Then, we use the cross-correlation function to calculate the delay between each frame of the right and left channel signals. The delay of the k-th is given by: $Delay(k) = |arg_{max}\{xcorr(S_{inR}(k), S_{inL}(k))\}|$. As the delay of turning head from left to right and the delay of turning head from right to left are different, we take the absolute value of $Delay(k)$ to remove the effect of the body movement's direction. Finally, we use the zero-crossing rates of dual channels, the correlation coefficient of dual channels, the mean of the delay profiling, and the standard deviation of the delay profiling to discriminate different micro movements.

(iv) Macro Movement Recognition: Limb Movement or Body Rollover? The cumulative power of two macro movements is shown in Fig. 14. We can observe that the cumulative power of body rollover is higher than limb movement. Based on Eq. 5, the vibration induced by limb movement needs to travel the whole body to the ears, which causes severe signal attenuation during propagation. The body rollover involves in limbs, head, and trunk. Although the vibrations generated by the limbs are still attenuated, the vibrations generated by the head and trunk can reach the ears with a little attenuation, which results in higher power in the spectrum. Thus, we use Q_0, Q_1, Q_2, Q_3, and Q_4 of the cumulative spectral power as features to discriminate the limb movement and body rollover.



(a) **In-ear snoring sound.** (b) **In-ear coughing sound.** (c) **In-ear somniloquy sound.**

Fig. 15. In-ear spectrograms of three sound activities.

5.2.4 ***Sound Activity Recognition.*** Due to the occlusion effect, in-ear body-conducted sounds are different from air-conducted sounds. Thus, we should analyze differences in in-ear sounds among different sound activities and then extract representative in-ear acoustic features for sound activity recognition.

(i) Behavioral Feature Extraction. The sound activities during sleep are continuous events that have correlations in time dimension. Snoring events have more distinct periodic patterns (*i.e.*, as periodic as breathing) over long periods of time [69], while coughing is rapid over a short period of time. Thus, as for a segmented event with start time point T_{start} and end time point T_{end}, we extract the upper RMS envelope of sounds within $[T_{start} - \Delta T, T_{end} + \Delta T]$, where ΔT denotes the time offset and is set as 8 s. Then we calculate the auto-correlation function of the extracted envelope. The location of the first peak of the estimated auto-correlation function is selected as τ, which is used as the periodicity feature.

(ii) Spectral Feature Extraction. Fig. 15 illustrates the spectrograms of three sound activities. We can find that the distribution of energy is different. For example, as for the snoring sound, the spectral power is mainly distributed below 1000 Hz, while the cough sound and somniloquy sound have a wider frequency band. In addition, the snoring sound and somniloquy sound have more obvious harmonic patterns (*i.e.*, white box) compared with the cough. In particular, the harmonic structures within [50, 500] Hz of three sound activities are different. Therefore, we select the location of the first five peaks of cumulative spectral power as $F_{harm} = \{F_{peak1}, F_{peak2}, F_{peak3}, F_{peak4}, F_{peak5}\}$. Then we use the standard deviation and mean of F_{harm} as harmonic-structure features for sound activity discrimination. We divide the frequency band ranging from [0, 2000] Hz into 4 non-overlapping sub-bands. For each sub-band, we calculate the energy ratio of this sub-band to the overall band. In addition, as for frequency components of each sub-band, we calculate the skewness and kurtosis as the measures of distribution. We also extract Mel-Scale Frequency Cepstral Coefficients (MFCCs) to represent the formant-related feature. In order to maintain the consistency of the feature dimension, we extract MFCCs from the whole event instead of each frame of the event. After feature extraction, we feed behavioral features and in-ear spectral features into a multi-class SVM for determination.

5.3 Physiological Activity Estimation

EarSleep aims to extract heartbeat and breathing waveforms that represent variations of physiological parameters in different sleep stages. The flow chart of physiological activity estimation is shown in Fig. 16. As introduced in Sec. 3.3, breathing-induced in-ear sounds and heartbeat-induced in-ear sounds are distributed below 3 KHz and 100 Hz, respectively. Therefore, we first downsample pre-processed audio signals to 8 KHz and 1 KHz respectively, and then use them as inputs for the breathing estimation module and heartbeat estimation module, respectively.

5.3.1 ***Heartbeat Estimation.*** The heartbeat-induced in-ear body sound is mainly distributed in $[0, 50]$ Hz. Thus, we first use a low-pass filter with a cutoff frequency of 50 Hz to keep heartbeat components. Fig. 17 shows

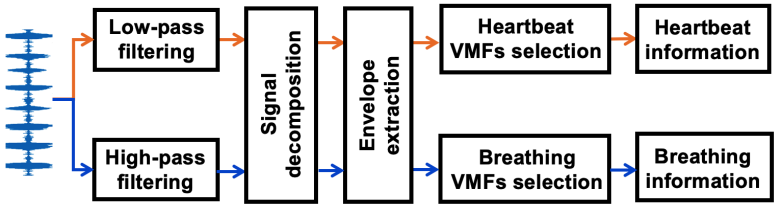


Fig. 16. Chart flow of physiological activity estimation. Input audio samples with a duration of 60 seconds are filtered and processed separately for breathing and heartbeat extraction.

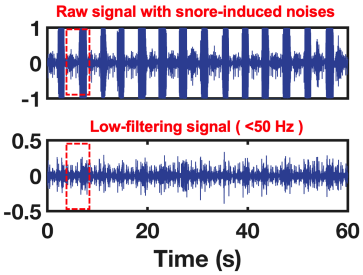


Fig. 17. Raw audio signal with snore noises (top). Low-pass filtering signal (bottom).

a 60-s in-ear sound sequence with periodic snore noises (shown in the red dotted box). However, after low-pass filtering, residual snore noises still interfere with the original waveform of heartbeat sounds. Next, we will detail how to extract heartbeat waveform from noisy in-ear sounds step by step.

(i) Signal Decomposition. The noise of body movement and sound activity seriously interferes with the heartbeat acoustic signal. We adopt the Variational Mode Decomposition (VMD) [15] to decompose noisy signals into multiple narrow-band modes (also called variational mode functions, VMF) with respective center frequencies. However, the performance of VMD mainly depends on the number of decomposition layers. If the number of decomposition layers is unreasonable, it may lead to mode mixing (*i.e.*, under decomposition) or induce the loss of useful components (*i.e.*, over decomposition). In our work, we use energy ratio E_r and dominant frequency spacing F_{bin} to determine the optimal k. Specifically, E_r is the indicator of under decomposition and can be expressed as:

$$E_r(k) = \frac{\left\| \sum_{i=1}^{k} vmf_i \right\|^2}{\|x\|^2} \quad k = 2, 3, 4, 5, 6. \tag{7}$$

where x, vmf_i, and k denote the input audio signals, the i-th variational mode function, and the number of modes, respectively. As the increase of k, $E_r(k)$ becomes larger gradually. It indicates that more useful information of the original signal can be kept in the reconstructed signal. F_{bin} is the indicator of over-decomposition, which is defined as the dominant frequency spacing between neighboring modes. In each iteration process, once F_{bin} is too close, it indicates that k is too large for decomposition. After the validation of 1000 one-minute heartbeat sounds with different noises, we find that $k = 4$ is sufficient for heartbeat waveform decomposition. Although other parameters, *e.g.*, penalty parameter and Lagrangian multiplier, are also related to the performance of decomposition, they have less contribution to the final result than k. In our work, we set them to the same default as [15].

(ii) SG-based Envelope Extraction. Fig. 18 shows four VMFs of in-ear acoustic signals. We can observe that there are non-stationary noises in each VMF. Thus, we adopt the Savitzky-Golay filter [58] to smooth each VMF to get a smoothed envelope. The Savitzky-Golay filter fits samples in each window using the quadratic polynomial, which can keep the shape and width of the origin signal unchanged while denoising. During sleep, the heartbeat intervals of adults range from [0.7, 1.2] seconds [48]. Therefore, the length of the filtering window is set as the minimum heartbeat interval and the sliding step is set as half of the length of the filtering window.

(iii) Heartbeat-related VMFs Selection. After smoothing the envelope of each VMF, we should assess which VMF has more heartbeat components and fewer interference components. Fig. 19(a) and Fig. 19(b) illustrate frequency analysis of smoothed VMF-2 and smoothed VMF-4, respectively. We can observe that amplitude of the snore component is larger than that of the heartbeat component in smoothed VMF-2. It indicates that smoothed

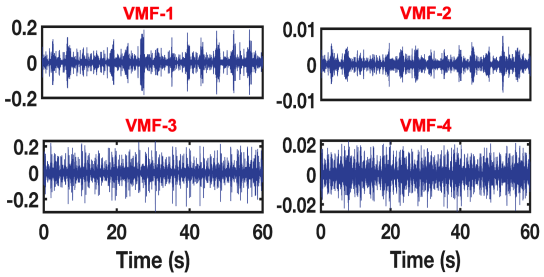
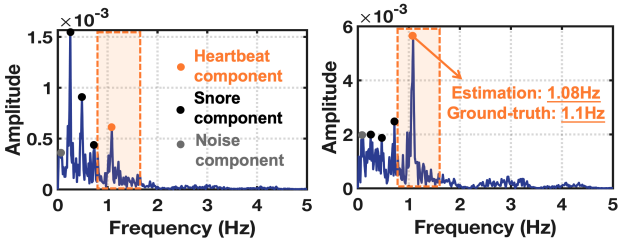


Fig. 18. Variational mode functions.



(a) FFT of VMF-2. (b) FFT of VMF-4.

Fig. 19. Frequency analysis of VMFs.

VMF-2 still contains more snore noises. On the contrary, the heartbeat frequency is dominant in smoothed VMF-4 which can be used for heartbeat information extraction.

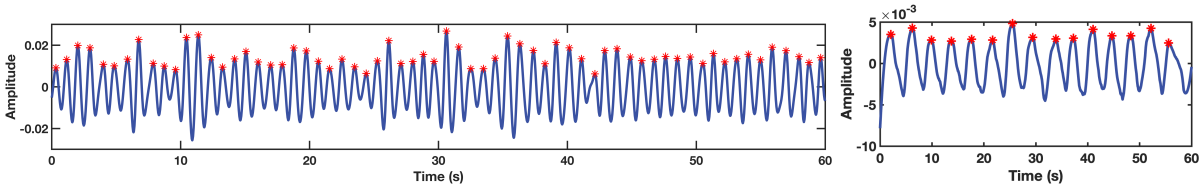
As for the optimal VMF, its dominant frequencies with high amplitude should be distributed in the range of 0.7-1.6 Hz (*i.e.*, the heartbeat frequency range of an adult). In particular, the heartbeat fluctuates slightly over time due to heart rate variability or other body movements' impact, which can result in the presence of one or more dominant frequencies. Therefore, based on the frequency-domain characteristics, we design two indicators, named heartbeat-to-noise ratio (HNR) and periodicity-to-noise ratio (PNR), to measure the ratio of heartbeat information in each VMF. HNR is the ratio of heartbeat components' amplitude to the overall amplitude, which indicates the contribution of heartbeat components. PNR indicates the periodicity of the heartbeat components. After the FFT operation, the spectral power is defined as $H = \{A(i)|0.8 <= f(i) <= 2\}$. The set of extreme points of H is defined as P. HNR and PNR are given by:

$$HNR = \frac{\sum H}{\sum\{A(i)|0 < f(i) <= 3\}}, \quad PNR = \frac{\sum\{P(i)|P(i) > \mu * H_{max}\}}{\sum H}. \tag{8}$$

where f, A, and H_{max} denote the frequency bin, the spectral power, and the maximum value of H. Since the heartbeat varies over time, it may have multiple frequencies. Only selecting the maximum value of H to participate in PNR computing is not enough. Thus, we select extreme points over $\mu * H_{max}$ from H to calculate the PNR, where μ is set as 0.8. The smoothed VMF with the highest $HNR + PNR$ is determined as the optimal VMF.

(v) Heartbeat Information Extraction. During sleep, some intensive body movements (*e.g.*, body rollover) may cause a rapid heart rate. The heart rate is relatively stable for a short period. Thus, we split the optimal VMF into several short-time segments with T_s seconds. As for each segment, we perform the FFT operation and extract the frequency component F_{heart} with the highest energy as the heartbeat rate. The number of heartbeats N_{heart} can be calculated by $N_{heart} = F_{heart} * T_s$. Then, a peak identification method is adopted to extract heartbeat intervals from the heartbeat waveform. The minimum heartbeat interval T_{min} between adjacent heartbeat peaks is set as 0.7 s [48]. Fig. 20(a) shows the heartbeat waveform with 60 seconds and estimated peaks, demonstrating that our system can accurately extract heartbeat waveform from noisy in-ear audio signals.

5.3.2 ***Breathing Estimation.*** The breathing is mainly polluted by body movements and heartbeat noises. When a user coughs or talks while sleeping, we consider that the user is not breathing. Snoring can be identified as a special type of breathing. To eliminate more noise components and keep more breath components, we first use the high-pass filter with a cutoff frequency of 100 Hz to remove the heartbeat noises. However, residual noises of body movements still interfere with the breathing waveform. Thus, we further take steps that are similar to heartbeat information estimation, to estimate the breathing-related information. The periodicity of breathing is not the same as the periodicity of heartbeat, so we need to fine-tune the methods introduced in Sec. 5.3.1.



(a) Heartbeat waveform. (b) Breathing waveform.

Fig. 20. Extracted heartbeat waveform (left) and breathing waveform (right).

Due to the wider frequency range of in-ear sounds of breathing, the low-frequency components of body movements can be divided from the high-frequency components of breathing. The levels of decomposition can be determined according to Eq. 7 and the dominant frequency spacing between neighboring modes, which can be set as 5. Then we use the Savitzky-Golay filter with a sliding window of 2 seconds and a sliding step of 1 second to smooth the noisy VMF. A normal person breathes 16-20 times per minute during sleep. Then, we calculate the energy set of breathing components as $H = \{A(i)|0.2 <= f(i) <= 0.4\}$. Then we define breath-to-noise ratio (BNR) as we define HNR in Sec. 5.3.1. Based on BNR and PNR, we select the VMF that contains the most breathing information. Compared with the heartbeat, the periodicity of breathing is more stable than the heartbeat. Thus, we can directly perform FFT operations to the optimal VMF to extract the frequency component F_{breath} with the highest energy. Then, the number of breaths is calculated by $F_{breath} * 60$. The peak identification method with $T_{min} = 2.5$ seconds is used for the breathing estimation. Fig. 20(b) illustrates the final extracted breathing waveform.

5.4 Sleep Stage Detection

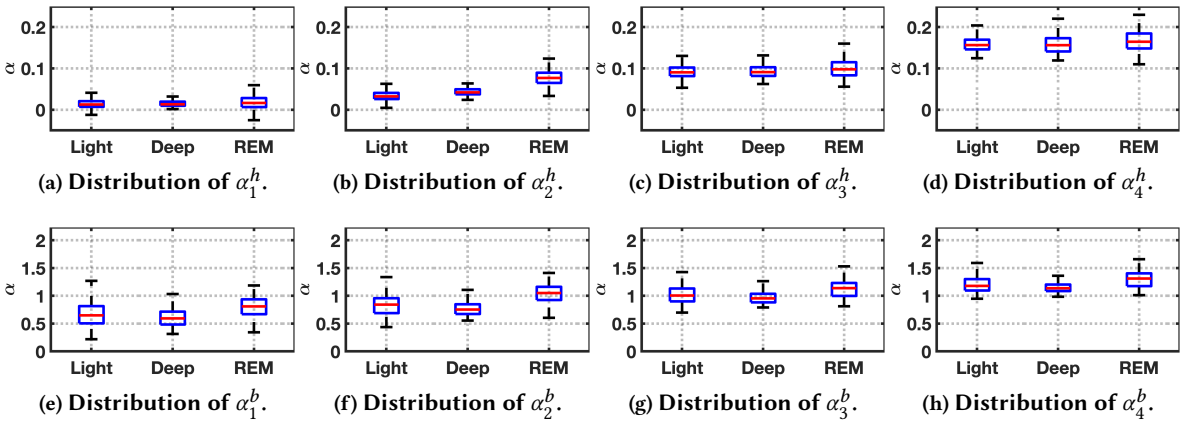
5.4.1 ***Sleep-related Features Extraction.*** Based on the above sleep events, we extract corresponding sleep-related features for sleep stage detection.

(i) Actigraphy Features. When the sleep becomes deeper, body movements become lighter and less frequent [31, 73]. In a sleep audio segment x with the length of N, we derive the number of body movements Num_{body} as the indicator of body movement's frequency. In addition, we also extract the amplitude ratio of body movements (denoted by R_{body}) as the intensity level of movements. Specifically, R_{body} is expressed as:

$$R_{body} = (\sum_{i=1}^{L_1} x_1(i)^2 + ... + \sum_{i=1}^{L_{Num_{body}}} x_{Num_{body}}(i)^2)/(\sum_{i=1}^{N} x(i)^2) \tag{9}$$

where L represents the length of a body movement event. In addition, the use of the amplitude ratio of various body movements instead of the direct sum of amplitudes can help eliminate differences between different individuals.

(ii) Sound Activity Features. Abnormal sound activities indicate the sleep stage changes. For example, a severe cough may cause people to wake up and the somniloquy often occurs in REM sleep stage. However, only small significant differences in lighter sound activities between different sleep stages. Therefore, we mainly focus on loud sound activities which are correlated with sleep stages [13, 32]. If the intensity of a sound activity is above a fixed threshold, it is determined as a louder sound activity. In addition, sound activity occurs at different frequencies in different sleep stages. For example, a higher frequency of snoring occurs in the deep sleep stage, compared with the REM and light sleep [32]. Thus, we count the number of occurrences of these three sound activities separately, which are denoted as $Num_{snore}, Num_{cough}, Num_{somn}$. Then, we also calculate the duration of these loud sound activities T_{sound} as the supplementary feature.



(a) Distribution of α_1^h. (b) Distribution of α_2^h. (c) Distribution of α_3^h. (d) Distribution of α_4^h.

(e) Distribution of α_1^b. (f) Distribution of α_2^b. (g) Distribution of α_3^b. (h) Distribution of α_4^b.

Fig. 21. Fluctuation scaling exponents with the n-th order polynomial fit of heartbeat waveform (top row) and breathing waveform (bottom row).

(iii) Physiological Activity Features. Previous studies [5, 31, 35] have explored the high correlation between variations in physiological activities and sleep stages. Next, several physiological activity features are extracted, namely waveform statistical features, correlation features, and CPC features. Before feature extraction, both heartbeat and respiratory waveforms are resampled to the same frequency.

(a) Waveform Statistical Features. However, directly using the heartbeat and breathing rate as candidate features may result in individual differences. Therefore, as for the k-th sleep segment, we calculate the cumulative distribution function of estimated heartbeat rate and breathing rate over the whole sleep period to represent sleep-related features [49]. Then, we extract the standard deviation of heartbeat intervals (*i.e.*, std_{hi}) and breathing intervals (*i.e.*, std_{bi}) to represent Heart rate variability (HRV) and breathing rate variability (BRV) which have been reported as effective indicators across different sleep stages [5, 31]. Furthermore, it has been found that the depths of breathing and heartbeat are more irregular during rapid-eye-movement (REM) sleep than during (NREM) sleep [35]. Thus, we also calculate the standard deviation of heartbeat's waveform amplitudes std_{ha} and breath's waveform amplitudes std_{ba} as the waveform-based features.

(b) Correlation Features. In different sleep stages, the breathing and heartbeat time series exhibit characteristic patterns of scaling behaviors [48, 59], which indicates that the long-time correlation properties of time series are potential indicators for identifying sleep stages. We adopt the detrended fluctuation analysis (DFA) method [72] which can eliminate the impact of non-stationary noises, to study the scaling behaviors of breathing and heartbeat time series. The fluctuation scaling exponents of heartbeat and breathing time series are denoted by α_n^h and α_n^b, respectively, where n denotes the order of polynomial fitting. As shown in Fig. 21, α_n^h and α_n^b show greater differences between three sleep stages. Specifically, the distributions of α_2^h and α_2^b are more discriminate across sleep stages, which are more suitable for sleep stage identification.

(c) Acoustic Cardiopulmonary Coupling Features. Cardiopulmonary coupling (CPC) measurements have been reported as one of the effective approaches for sleep quality estimation [40, 65], which represents the correlation between heartbeat information and breathing information. As for the heartbeat time series $x_h(t)$ and breathing time series $x_b(t)$, the acoustic cardiopulmonary coupling F_{ACPC} measurement can be derived via the product of cross-spectral power $\Gamma(H, B)$ and magnitude squared coherence $\Psi(H, B)$, *i.e.*, $F_{ACPC} = \Gamma(H, B) * \Psi(H, B)$.

$$\Gamma(H, B) = A_H A_B e^{-j(\Phi_H + \Phi_B)} \quad \Psi(H, B) = \frac{\Gamma(H, B)^2}{(A_H e^{-j\Phi_H})^2 (A_B e^{-j\Phi_B})^2}$$

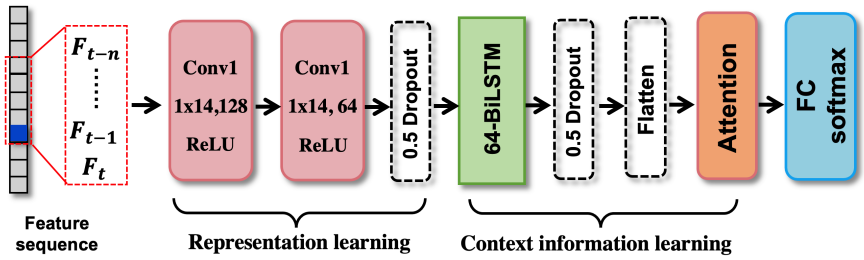


Fig. 22. Structure of deep learning model for sleep stage detection.

where A_H, and Φ_H respectively denote the amplitude spectrum and phase spectrum of the heartbeat series, and A_B and Φ_B respectively denote the amplitude and phase of the breathing series. Then we calculate the ratio of the sum of F_{ACPC} in [0.01, 0.1] Hz to the sum of F_{ACPC} in [0, 0.4] Hz, and the ratio of the sum of F_{ACPC} in [0.1, 0.4] Hz to the sum of F_{ACPC} in [0, 0.4] Hz, as the acoustic cardiopulmonary coupling features.

5.4.2 ***Sleep Stage Inference Model.*** Sleep is a cycling process that consists of various sleep stages (*i.e.*light sleep, deep sleep, and REM sleep). Prior works [19, 76] have reported the sleep stages follow a predictable transition pattern (*e.g.*, light sleep → deep sleep → light sleep → REM sleep), which indicates that there are dependencies or connections between sleep stages before and after the current moment. In our work, we design a BiLSTM-based model to classify three-class sleep stages (*i.e.*, REM, light sleep, and deep sleep). The structure of the sleep stage recognition model is shown in Fig. 22. As for the current time period t, we use features of the current time period F_t and features of the previous time periods $[F_{t-1}, F_{t-2}, ..., F_{t-n}]$ to infer the sleep state of the current time period. Then the feature sequence $[F_t, F_{t-1}, ..., F_{t-n}]$ is firstly fed into the CNN block which consists of two convolutional layers with rectified linear unit (ReLU) activation and a dropout layer. Each convolutional layer extracts high-level representation from the feature sequence. The dropout layer helps prevent overfitting and improve the generalization of the model. Then, the BiLSTM layer with 64 hidden nodes is used for context information analysis, which can learn sequential dependencies in both directions (*i.e.*, forward and backward). In addition, we add the attention mechanism behind the BiLSTM layer to learn the transition patterns of sleep stages. The attention layer assigns different weights to the output of the BiLSTM layer. Finally, a fully connected layer with softmax is added to classify sleep stages.

6 PERFORMANCE EVALUATION

6.1 Experimental Setup

6.1.1 ***Participants.*** We recruited 18 participants (12 males and 6 females) from our institute to participate in the evaluation. These participants ranged in age from 21 to 32 years old. All participants do not suffer from severe diseases such as cardiovascular disease, high blood pressure, and respiratory diseases. In particular, three participants who had just recovered from COVID-19 for a week to a month and still suffered from coughing. Two participants suffer from severe insomnia and often require supplemental medication to fall asleep. Before conducting research, we obtained voluntary informed consent from research participants [2].

6.1.2 ***Prototype Implementation.*** We embed low-cost and small-size microphones into soft earplugs which are shown in Fig. 23(a). The sampling frequency of the in-ear microphone is 44.1 KHz. We use the Boya dual-channel

[2]While our institute has not yet established the Institutional Review Board (IRB), as an alternative mechanism, the entire research process is subject to the supervision of the Institution's Academic Committee (https://en.cs.ustc.edu.cn/23588/list.htm) and the local regulations, ensuring the ethical conduct of our study and the protection of participants' rights. Before conducting research, we have obtained voluntary informed consent from research participants.

(a) Hardware implementation. (b) Experimental environment.

Fig. 23. Hardware prototype implementation and experimental environment.

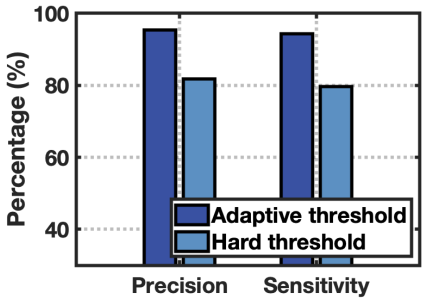


Fig. 24. Evaluation of the event segmentation method.

Table 1. Evaluation of the event type classification method.

		Estimation		Recall
		Body	Sound	
Ground Truth	Body	1534	46	97.09%
	Sound	27	1879	98.58%
Precision		98.27%	97.61%	

recording interface to connect earbuds to a Lenovo Work Station that is equipped with two Inter Xeon Gold 6226R@2.9 GHz CPUs and 128 GB memory. We use MATLAB-R2021B to process raw audio signals and extract features. The model of sleep stage inference is constructed based on Tensorflow *v1.9.0* and Keras *v2.2.0*.

6.1.3 ***Experimental Data and Ground Truth Collection.*** We deploy our system in a quiet bedroom which is shown in Fig. 23(b). We deploy a Xiaomi home camera Pro with infrared night vision and dual microphones to detect body movements and sound activities during sleep. Before the experiment, the participant wears earbuds in the most comfortable way. In addition, we also ask the participant to wear a Polar H9 heart rate belt and a HKH-11L breathing belt on his/her chest which can provide the ground truth heart rate and respiration rate. These consumer-grade devices are well-known for their accuracy in physiological activity monitoring and have been widely adopted in various studies for providing reliable ground truth [8, 62, 64, 70]. Since professional medical devices such as PSG are expensive and complicated to operate, we leverage the result of MeetSleep [44] which is an EEG-based sleep monitoring device, as the ground truth of sleep stages. EEG signals can directly reflect the brain activities, which can be considered an effective way for sleep analysis [39]. After wearing the devices, each participant sleeps about 6-8 hours per night during his/her normal sleep schedule. Each participant contributes 2-3 nocturnal sleep data. In total, we collect 243.9 GB of sleep audio data for 48 nights.

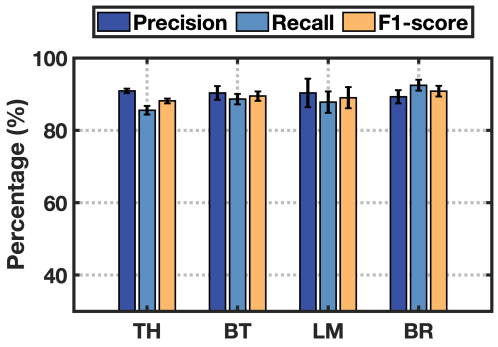


Fig. 25. Evaluation of body movement recognition.

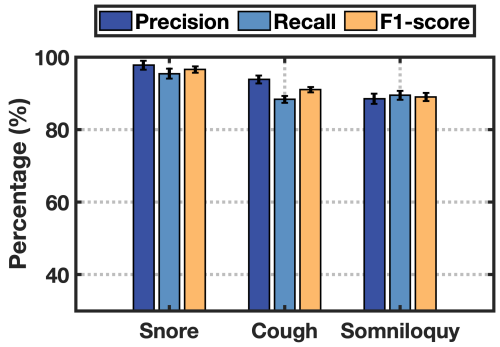


Fig. 26. Evaluation of sound activity recognition.

6.2 Evaluation of Event Segmentation and Event Type Classification.

(i) Effectiveness of Event Segmentation. We adopt event-based metrics that are commonly used in sound event detection tasks [30, 41]. The estimated temporal location of an event consists of the start location denoted by T_s and the end location denoted by T_e. The ground truth temporal location of an event can be denoted by $[G_s, G_e]$. For each audio segment, we use a sliding hamming window with a length of 0.2 seconds and a step with a length of 0.1 seconds to split the acoustic signals into N frames for envelope extraction. If the estimated temporal location overlaps with the ground truth temporal location, we think that the event is correctly detected. In addition, for the start temporal location and the end temporal location, we allow a tolerance of ± 200 ms (*i.e.*, two frame-shift of 100 ms [41]). If the T_s is in G_s ± 200 ms and the T_e is in G_e ± 200 ms, the event is correctly detected. The true positive instance(TP), false positive instance(FP), and false negative instance(FN) are defined as similar to [41]. The precision is denoted by $P = TP/(TP+FP)$ and the sensitivity is denoted by $P = TP/(TP+FN)$. Fig. 24 shows the event segmentation results. Compared with the hard threshold, the precision and sensitivity of the adaptive threshold are increased by 13.6% and 14.4%, respectively.

(ii) Effectiveness of Event Type Classification. In the data collection, we collected and labeled 1580 instances of body movements (368 instances of turning head, 236 instances of body trembling, 436 instances of limb movement, and 540 instances of body rollover) and 1906 instances of sound activities (1485 instances of snore, 305 instances of cough, and 116 instances of somniloquy). Then, we use the event type classification method to determine whether each instance is a body movement or a sound activity. Tab. 1 shows the event type classification results. We can find that our system can effectively identify each type of sleep event with high precision and recall.

6.3 Evaluation of Body Movement Recognition.

As for 4 common body movements related to sleep disorders and movement disorders, we give a comprehensive evaluation of the performance of body movement recognition. The automatic behavior recognition function of the camera can help us label body movements. For convenience, we refer to turning head, body trembling, limb movement, and body rollover as TH, BT, LM, and BR, respectively. As for each participant, we use 10-fold cross-validation to evaluate our system. The overall average precision, average recall, and average F1-score among 18 participants are shown in Fig. 25. The recognition performance of macro movements (*i.e.*, limb movement and body rollover) is better than that of micro movements. That is because the energy patterns of body rollover and limb movement are obviously distinguishable. Some limb movements with short durations are easily misclassified as micro movements, which causes low recall of limb movement recognition. As for micro movements, the recall

of turning head is about 86%, which is lower than other movements. It is because some participants continuously turn their heads during sleep, causing the system to misjudge these movements as macro movements.

Table 2. Body movement recognition of Participant B

Participant B				
	TH	BT	LM	BR
Presision	86.67%	66.04%	62.26%	93.75%
Recall	69.64%	85.37%	91.67%	60.00%

Table 3. Body movement recognition of Participant C

Participant C				
	TH	BT	LM	BR
Presision	89.58%	78.79%	94.00%	90.00%
Recall	84.31%	86.67%	92.16%	91.84%

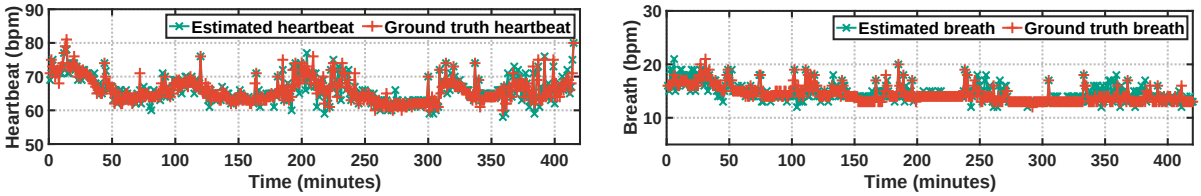
Since we classify different body movements based on the physical characteristics of the body, the diversity of human body structures is an important factor that may affect the results of recognition. We randomly select three participants (*i.e.*, *A, B,* and *C*) with different heights and weights to explore the impact of individual differences. We use the data of *Participant A* (179 cm/90 Kg) as the training set. The data of *Participant B* (165 cm/65 Kg) and the data of *Participant C* (184 cm/94 Kg) are used as the testing set. Tab. 2 and Tab. 3 show the classification results of *Participant B* and *Participant C*, respectively. EarSleep yields the accurate recognition performance of *Participant C* with an average precision of 88.09% and an average recall of 88.74% among four movements, outperforming the recognition performance of *Participant B* by 10.91% and 12.07%, respectively. Although the above results indicate that individual differences in body movement recognition can not be ignored, the performance of body movement recognition among individuals with similar body sizes still meets the needs of the usage of sleep detection. More discussion about individual differences is introduced in Sec. 7.

6.4 Evaluation of Sound Activity Recognition.

Among the total participants, there are 12 persons who perform snore, cough or somniloquy during sleep. Specifically, 9 (6 males and 3 females), 6 (4 males and 2 females), and 3 (2 males and 1 female) persons among them snored, coughed, and talked during their sleep, respectively. Then, we segment 1485 instances of snore, 305 instances of cough, and 116 instances of somniloquy. Considering the imbalance of the dataset, we also use 10-fold cross-validation to evaluate the performance of our system. Fig. 26 shows precision, recall, and recall of sound activity recognition. EarSleep can detect snores with an average precision of 97.05%, an average recall of 94.99%, and an average F1-score of 96.01%, which indicates that our system can effectively track snores during sleep. The worst classification result belongs to somniloquy, which can still achieve an average precision of 89.28%, an average recall of 91.30%, and an average F1-score of 90.28%, respectively. In addition, there are other sound activities such as bruxism that may occur during sleep. However, due to the limited number of participants in our experiments, we are unable to collect these related samples for analysis and detection. With the increase in the number and diversity of user groups, we believe our system can detect more sleep-related sound activities.

6.5 Evaluation of Physiological Activity Estimation.

6.5.1 ***Overall Performance.*** We evaluate the long-term performance of physiological activity estimation by calculating heartbeat and respiration rates in each minute of the night. We only use the left-ear channel acoustic samples to extract heartbeat and breathing rates for performance analysis. Fig. 27(a) and Fig. 27(b) show continuous heartbeat rate and breath rate estimation throughout the night, respectively. We observe that EarSleep can yield accurate heartbeat and breath rates with a mean absolute error of 1.61 bpm and 1.21 bpm, respectively. We also observe that there are potential outliers in Fig. 27. We speculate the potential reasons for the presence of potential outliers as follows: (1) the outliers are caused by poor contact between the heartbeat/breathing belts and the body surface due to some movements; (2) inherent variability/fluctuation patterns of heartbeat and breathing [6, 35]. Furthermore, we calculate the heartbeat and breath rates of all participants involved in our experiments. Fig. 28



(a) Heartbeat rate estimation.

(b) Breathing rate estimation.

Fig. 27. Heartbeat rate and breathing rate estimation throughout the night.

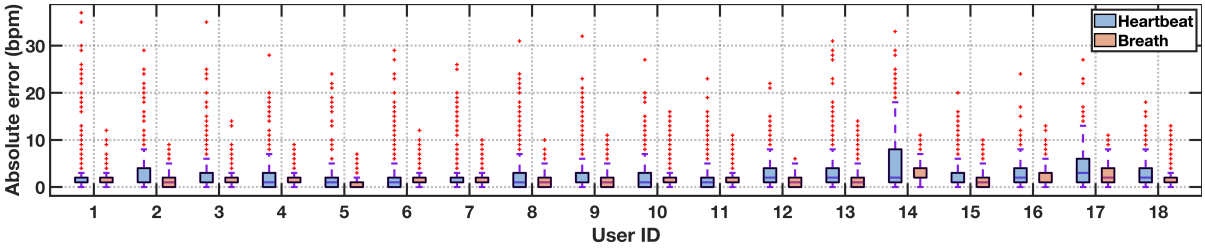


Fig. 28. Mean absolute errors (MAE) of heartbeat and breathing rate estimations among all participants.

shows the distribution of absolute errors of heartbeat rate and breath rate estimation. The mean absolute errors of heartbeat and breath among all participants range from [2.16, 5.24] bpm and [1.03, 3.29] bpm, respectively.

6.5.2 ***Analysis of Special Case.*** There are larger fluctuations in the heartbeat and breath patterns of some participants (*i.e.*, *Participant 14* and *Participant 17*). *Participant 14* suffers from the heart rhythm problem (*i.e.*, heart arrhythmias) according to his medical history and diagnosis, whose heart may sometimes beat too quickly during sleep. Therefore, rapid changes in the heartbeat and breath rates over a period of time make our estimation inaccurate. In addition, *Participant 17* has just recovered from COVID-19 one week and wakes up frequently with a severe cough, which may cause abnormal heartbeat and breathing and variability patterns.

6.5.3 ***Consistency Analysis with Ground Truth.*** To further analyze the differences in physiological activity measurements between EarSleep and ground truth, we randomly select data of ten nights and leverage a Bland-Altman plot to visualize the consistency between two measurement techniques. Specifically, in Fig. 29, the average difference (*i.e.*, horizontal blue line) between heartbeat ground truths and estimated heartbeat rates is 0.11 bpm. About 96.54% of points are distributed in the confidence interval (*i.e.*, the area between two dotted red lines) of the average difference, which indicates that heartbeat estimations of EarSleep are highly consistent with heartbeat ground truths. Similarly, as shown in Fig. 30, the average difference between the two measurement techniques is -0.87 bpm and about 97.13% points are distributed in the confidence interval, which denotes that the two techniques of breathing measurements are also in agreement.

6.5.4 ***Impact of Acoustic Channels.*** We also use right-ear channel acoustic samples to evaluate the performance of physiological activity estimation. The scheme of dual-channel fusion is to average the results of the right-ear channel and the left-ear channel. The distributions of absolute errors between estimations and ground truths are shown in Fig. 31. Overall, there are only subtle differences between the left-ear channel and the right-ear channel in heartbeat and breathing estimation. The mean absolute errors of the three schemes (*i.e.*, left channel, right channel, and dual channel) are 3.24 bpm, 4.29 bpm, and 3.60 bpm for heartbeat estimation, respectively. We think that is because the heart is on the left side of the human body, resulting in a stronger

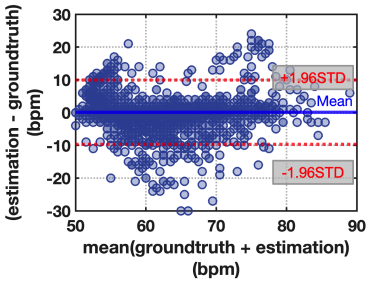


Fig. 29. Bland Altman plot of heartbeat rate estimation.

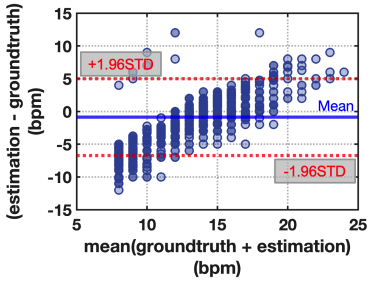


Fig. 30. Bland Altman plot of breathing rate estimation.

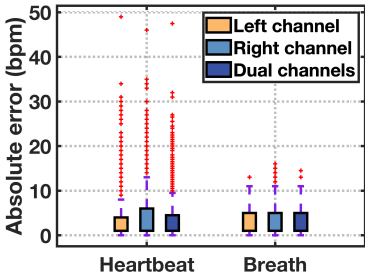


Fig. 31. Impact of acoustic channel on physiological activity estimation.

Table 4. Overall performance of sleep stage detection.

	REM sleep			Light sleep			Deep sleep		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
w/o Attention	70.74%	72.31%	71.52%	68.21%	66.61%	67.40%	62.64%	58.47%	60.48%
Attention	74.13%	75.65%	74.88%	72.17%	66.25%	69.08%	68.17%	61.39%	64.60%

acoustic response of the left-ear channel than that of the right-ear channel. The mean absolute errors of the three schemes are 2.98 bpm, 2.99 bpm, and 2.88 bpm for breathing estimation, respectively. Thus, only a single acoustic channel is sufficient for physiological activity estimation.

6.6 Evaluation of Sleep Stage Detection.

6.6.1 ***Overall Performance.*** Collected acoustic samples were split into 1-minute segments and manually labeled. We use 10-fold cross-validation to evaluate the overall performance of sleep stage detection/inference. Table 4 illustrates the results of sleep stage detection with and without attention mechanism assistance, where we can obviously observe EarSleep achieves an average precision of 71.49%, an average recall of 67.76%, and an average F1-score of 69.52% among three sleep stages with attention mechanism assistance, significantly outperforming no attention mechanism assistance method by 4.29%, 1.97%, and 3.05%, respectively. That is because the attention mechanism learns the context relationship between time series and adds the transition restriction on the sleep stages during prediction. We can also observe that EarSleep has a better detection performance on REM sleep, followed by light sleep and deep sleep. It indicates that sleep-related acoustic features we extract in Sec. 5.4.1 represent more characteristics of REM sleep. In our experiments, we also find that the changes in physiological activity of the same sleep stage also vary from person to person. For example, compared with other participants, a participant who exercises regularly shows less variation in heartbeat and respiration during different sleep states. It indicates that our system cannot totally eliminate individual differences. The purpose of our system is to provide users with a daily sleep quality reference via a portable way.

6.6.2 ***Feature Validation.*** We select one night's sleep data from each participant to evaluate the contribution of features to the final result. In our work, when detecting different sleep stages, we take actigraphy features, sound activity features, and physiological features into account. Actigraphy features and sound activity features have been reported as effective features for sleep stage detection in previous solutions [12, 19]. To assess the validity of the physiological features, we adopt the features with and without physiological features assistance for sleep stage detection. As shown in Table 5, with the assistance of physiological features, our system improves

Table 5. Performance of sleep stage detection using different features.

	REM sleep			Light sleep			Deep sleep		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
w/o Physiological Feature	49.31%	47.29%	48.28%	41.28%	43.67%	42.44%	40.23%	42.47%	41.32%
with Physiological Feature	75.31%	74.23%	74.77%	70.69%	66.19%	68.37%	63.74%	65.33%	64.53%

Table 6. Performance comparison with other solutions.

	REM sleep			Light sleep			Deep sleep		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Sleep Hunter	60.91%	57.66%	59.24%	54.12%	50.23%	52.10%	49.48%	45.78%	47.56%
SleepGuard	67.79%	64.27%	65.98%	60.78%	56.42%	58.52%	62.84%	56.32%	59.40%
EarSleep	74.21%	77.37%	75.76%	72.39%	65.32%	68.67%	66.17%	62.29%	64.17%

Table 7. Latency evaluation of EarSleep.

	Pre-processing	Physical activity	Physiological activity	Sleep stage detection
Average latency (s)	0.16±0.01	11.33±1.25	27.69±0.99	3.44±0.03

the average precision by about 26%, average recall by about 24%, and average F1-score by about 25%. It indicates that physiological features play an important role in sleep stage detection.

6.6.3 ***Comparison with Other Solutions.*** Although existing works [13, 50, 53, 61] have studied portable sleep monitoring using ubiquitous sensors on commercial smart devices, most of them focus on detecting limited sleep events and are unable to detect sleep stages. Thus, we select Sleep Hunter (mobile phone-based) [19] and SleepGuard (smartwatch-based) [12] as our baselines, which are considered representative solutions in the field with sleep stage detection. The comparison results are shown in Table 6. We can observe that our system outperforms other methods, benefiting from our fine-grained features. In particular, compared to SleepGuard and Sleep Hunter, EarSleep not only captures body movements and sound activities, but also captures physiological activities containing rich sleep-related information.

6.7 Latency Evaluation

Due to the extremely limited computing resources of current earbuds, our system runs on edge devices (*i.e.*, Lenovo Work Station), as described in the experimental setup. We divide our system into four parts and calculate the latency of each part. The first part is pre-processing (Sec. 5.1). The second part is event recognition which includes event segmentation (Sec. 5.2.1), coarse-grained event classification (Sec. 5.2.2), body movement recognition (Sec. 5.2.3), and sound activity recognition (Sec. 5.2.4). The third part is physiological activity estimation (Sec. 5.3). The last part is sleep stage detection (Sec. 5.4) which includes sleep-related features extraction and sleep stage inference. EarSleep first performs the pre-processing part, then performs the event recognition and physiological activity estimation in parallel, and finally executes sleep stage detection. Thus, the entire processing latency can be calculated by $T_{pre} + max(T_{event}, T_{phy}) + T_{sleep}$, where T_{pre}, T_{event}, T_{phy}, and T_{sleep} denote the latency of pre-processing, event recognition, physiological activity estimation, and sleep stage detection, respectively. The

average latency is shown in Table 7. Specifically, the part of physiological activity estimation accounts for the most latency due to signal decomposition and VMF selection Furthermore, we do a qualitative analysis of latency comparison between our work, SleepGuard, and Sleep Hunter. Sleep Hunter and SleepGuard utilize the outer microphones to capture the sound activities (*e.g.*, snoring, coughing, and talking in sleep), their latency in sound activity recognition is similar to ours by analyzing the acoustic signal processing methods they use. However, other sensors used in SleepGuard and Sleep Hunter have a lower sampling rate. For example, in addition to the microphone, the other four sensors used in SleepGuard (*i.e.*, accelerometer, gyroscope, microphone, light, and orientation sensor) all have sampling frequencies below 50 Hz, as mentioned in their article. Thus, compared with the high-sampling acoustic signals, non-acoustic signals captured by sensors with low sampling frequency require fewer resources to process, resulting in lower latency. Our goal is to provide continuous sleep monitoring and analysis throughout the night. Thus, the emphasis of our research is on the accuracy and reliability of long-term measurements and the real-time response is worthy of exploring in the future.

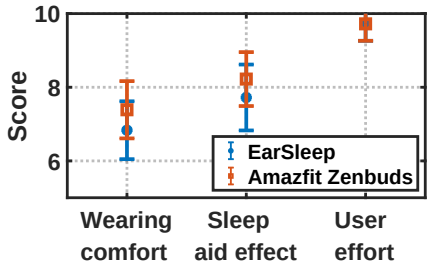


Fig. 32. User experience study.

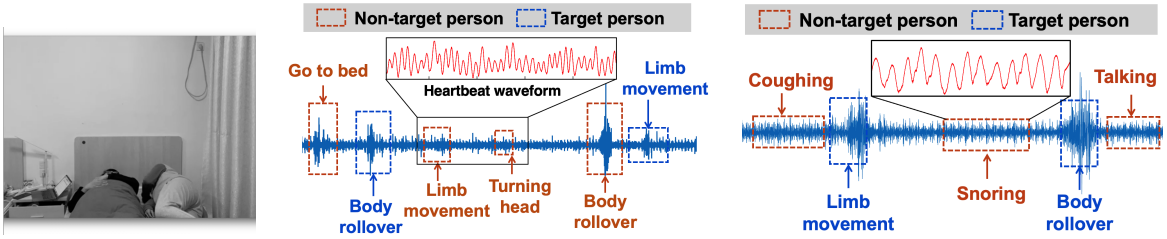
6.8 User Experience Study.

We compare the user experience of our system with Amazfit Zenbuds, commercial sleep earbuds that are equipped with biological sensors for sleep monitoring. We recruit 10 participants and ask them to fall asleep with EarSleep and Amazfit Zenbuds. In particular, we focus on three metrics for the user study – *wearing comfort*, *sleep aid effect*, and *user effort*. The *wearing comfort* refers to the pressure and pain induced by external devices during sleep. The *sleep aid effect* refers to the impact of external devices on sleep quality. The *user effort* refers to the complexity or difficulty of operation when the user uses the external device. Participants are required to complete a post-study survey with a 10-point Likert scale (from one to ten) after waking up every day. The higher score in each item means higher user experience and satisfaction. Fig. 32 shows the results of the user study. EarSleep could bring a comparable user experience with Amazfit Zenbuds. Both EarSleep and Amazfit Zenbuds adopt the design of silicone ear tips which can block out external noises at night and provide a quiet environment for sleep. In terms of wearing comfort, we find that the score of EarSleep is lower than Amazfit Zenbuds. Since the sizes of ear canals vary from person to person, the sleep earbuds that we prepare cannot perfectly fit every participant's ear canals. As a result, some participants find their ear canals uncomfortable. Nevertheless, this can be readily addressed using softer silicone material and earbud shape design fitting the structure of the ear canal [55].

7 DISCUSSION

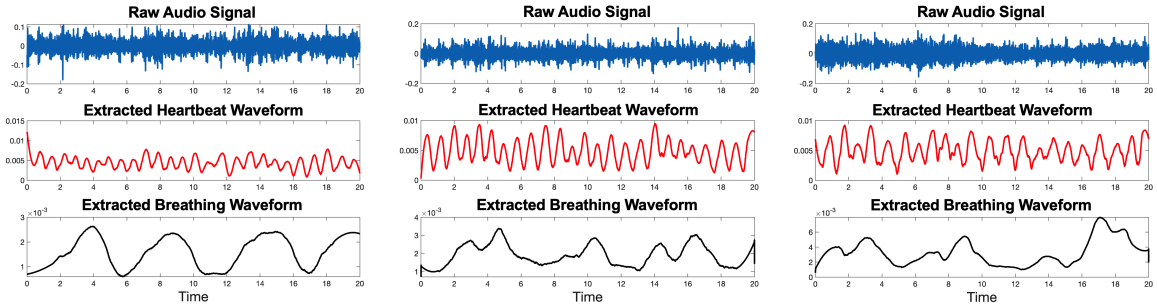
Our work aims to show the feasibility and availability of sleep monitoring via the in-ear acoustic modality. However, there is still a gap between our system and large-scale applications, where some practical issues should be addressed.

i) Usage in Multi-person Scenarios. While our work aims to provide individual sleep monitoring, we still understand the importance of usage in multi-person scenarios. We simulate a scenario where two persons lie on



(a) Illustration of a multi-person scenario. (b) Impact of body movement interference from non-target person. (c) Impact of sound activity interference from non-target person.

Fig. 33. The target person wears EarSleep and lies on the bed (left side). The non-target person sleeps with the target person together and performs body movements and sound activities. The blue dotted box and organ dotted box represent the acoustic response of the target person and that of the non-target person, respectively.



(a) Average noise level of 48 dB. (b) Average noise level of 55 dB. (c) Average noise level of 67 dB.

Fig. 34. Impact of ambient noise on the in-ear acoustic signal under three noise levels.

the same bed, as shown in Fig. 33(a). Next, we explore the impact of interference caused by the non-target person and the external environment on the target in-ear acoustic signals from three aspects.

Body movement interference from the non-target person. The non-target person naturally performs body movements. The in-ear acoustic signals collected from the target person are shown in Fig. 33(b). We find that only intensive body movements of the non-target person like body rollover can interfere with in-ear acoustic signals of the target person. Lighter body movements of the non-target person like limb movement and turning head have a subtle impact on the target person. For instance, as shown in the black box of Fig. 33(b), EarSleep can accurately extract the heartbeat waveform of the target person when the non-target person performs limb movement and turning head.

Sound activity interference from the non-target person. Similarly, Fig. 33(c) illustrates in-ear acoustic signals of the target person when the non-target person performs sound activities (*i.e.*, snoring, coughing, and talking). We can observe that sound activity interference has a negligible impact on the in-ear acoustic signals of the target person. For instance, EarSleep can still accurately extract the heartbeat waveform of the target person when the non-target person performs snoring. Earplugs on sleep earbuds can block the external noise into ear canals, making in-ear acoustic signals of the target person resistant to external sound interference.

Ambient noise interference from the external loudspeaker. Furthermore, we also explore the impact of ambient noise on in-ear acoustic signals. Specifically, we put a loudspeaker at the bedside table which is close (about 40 cm) to the target person. We play music at different volumes to simulate the ambient noise with different noise

levels. Considering that physiological activities (*i.e.*, heartbeat and respiration) are lighter than physical activities (*i.e.*, body movements and sound activities). Thus, we analyze the impact of ambient noise on the physiological activity estimation to initially explore the robustness of EarSleep in different noise levels. The results are shown in Fig. 34. We can clearly observe that EarSleep can accurately extract heartbeat and breathing waveforms even in the average ambient noise level of 55 dB. When the average ambient noise level achieves 67 dB, EarSleep is unable to extract accurate the heartbeat and breathing waveforms. However, according to recommendations of the World Health Organization, the comfortable sleeping environment noise level for humans is about 30-40 dB. Therefore, in real life, people also cannot fall asleep in such noisy environments.

ii) Privacy Concerns. The primary goal of our work is to design a portable sleep monitoring system based on acoustic signals. However, the signals may include sensitive information of the user that may involve privacy concerns. In particular, since EarSleep needs to process long-term in-ear audio signals for sleep monitoring, it may cause concerns about the leakage of important physical and physiological information hidden in audio signals. To reduce the risk of privacy leakage to a certain extent, in the principle of the Nyquist-Shannon sampling theorem, down-sampling could cause the loss of high-frequency signals that include more human-voice information. However, through the above analysis, the high-frequency information is an important feature for body movement recognition and sound activity recognition. While we already downsampled the signals from 44.1 KHz to 8 KHz, there might still be remaining information with privacy concerns. In our study, we have taken strict measures to protect privacy in the entire process of research. However, it is still challenging to strike a balance between performance and privacy at the current stage. Nevertheless, future work can focus on pushing the boundary of the high-frequency range to a lower value for improving privacy protection.

iii) Individual Difference. The ages of participants in our evaluation only range from 21-32 years old, which does not cover a wider age range. Older people's metabolism is slower than younger people's, resulting in different physiological patterns during sleep. Therefore, one limitation of our work is that we cannot provide sleep analysis of people with different age groups. In addition, some diseases also influence sleep quality as mentioned in Sec. 6.5.2. In practice, we are difficult to collect sufficient sleep data for these special cases, preventing us from providing the analysis of the impact of diseases. In the future, we will actively establish cooperation with medical institutions that provide more resources for us to complete this interesting and meaningful research topic.

iv) Sleep Music Playback over Earbuds. Some commercial sleep earbuds support sleep music play mode which helps people relax and fall asleep. When earbuds start to play sleep music (*e.g.*, white noises), in-ear sounds induced by various sleep activities are distorted. For example, the intensity of raw in-ear breathing sound during sleep usually ranges from 18-30 dB and the frequency of in-ear breathing sound is mainly distributed in 0-1500 Hz, while comfortable white noises are about 40-60 dB and the frequency range is mainly distributed in 0-3000 Hz. With the development of deep learning, DL-based methods (*e.g.*, adversarial learning) have been verified in many fields to achieve better performance than traditional methods. In the future, we may use deep learning to learn the differences between in-ear sounds and music noises and extract pure target signals. In addition, motion sensors have become a novel sensing approach to capture sound vibration [3, 22, 23]. These motion sensors are widely equipped on earbuds and may provide a complementary approach to remove the impact of sound playback.

v) Sleep Posture Detection. Sleep postures (*i.e.*, supine, prone, right lateral, and left lateral) determine the comfort of sleep. Improper sleep postures cause stress on the body, contributing to low-quality sleep and sleep-related diseases [11, 46]. It is important to provide sleep posture detection and tracking overnight which can help people understand the impact of sleep postures on personal sleep quality and body health. One common way to achieve sleep posture detection is to equip earbuds with motion sensors. However, we find an opportunity to achieve sleep posture detection using in-ear heartbeat sounds. Different sleep postures cause different stress on the heart and respiration tract, resulting in special variation patterns of in-ear sounds. We believe using in-ear acoustic signals to achieve sleep posture detection is challenging but worth exploring in the future.

8 RELATED WORK

8.1 Non-contact Approaches

i) Wireless-based. Many wireless-based approaches have been proposed to sense and understand sleep behaviors. The intuition behind these wireless-based approaches is to capture unique variation patterns of wireless signals induced by sleep activities. Based on these unique signal patterns, sleep events like breathing [34, 74, 76, 77, 79–82], heartbeat [20, 34, 75], sound activities[33], and sleep postures [78] can be detected. However, most of the above systems require a pair of wireless transceivers in the environment. Due to the high sensitivity of wireless signals to external environments, once the environment is changed, the detection accuracy may decrease. In addition, these solutions only detect several sleep events and are not sufficient to provide fine-grained sleep information.
ii) Smartphone-based. With the increase of built-in sensors in smartphones, numerous approaches use commercial smartphones to monitor sleep. Chang *et al.* [13] develop a mobile phone software, named *iSleep*, which processes sounds produced during sleep for sleep quality monitoring. Gu *et al.* [19] propose *Sleep hunter* which leverages multiple sensors equipped on smartphones, including the microphone, the light sensor, and the accelerometer, to provide fine-grained sleep detection. However, these smartphone-based approaches require the smartphone to be placed close to the user's head, which is not always practical. For example, large body movements of the user during sleep will change the smartphone's location.

8.2 Wearable-based (Contact) Approaches

With the development of embedded technology, more and more sensors are embedded in smart devices. Compared with non-contact sleep monitoring approaches, wearable-based approaches can capture more fine-grained physical sleep information due to contact with the human body and show higher robustness in different environments.
i) Body-worn. Most of the key physiological activities (*e.g.*, heartbeat and breathing) are performed in the human upper body. Therefore, body-worn approaches are proposed to achieve upper body motion and physiological activity detection for long-term sleep monitoring. The approaches proposed in [27, 60] integrate RFID into the fabric to achieve breathing, heartbeat, and upper-body motion detection overnight. *Phyjama* proposed in [28] leverages four resistive sensors and a tribolelectric sensor embedded into the pyjamas for physiological sensing. He *et al.* [25] develop a smart belt with accelerometer and pressure sensor to detect sleep-related events including vital signs and snore. However, the aforementioned body-worn approaches need additional sensors to sense sleep events, which are not directly deployed on commercial devices.
ii) Wrist-worn. Biomedical sensors(*e.g.*, PPG and ECG) have been embedded into smart wrist-worn devices to measure physiological information. However, these specialized biomedical sensors are expensive, prompting researchers to use ubiquitous and low-cost sensors (*e.g.*, microphone, accelerometer, light sensor) on wearable devices for sleep monitoring [12, 50, 61]. Chang *et al.* [12] exploit *SleepGuard* that uses various sensors on the smartwatch to capture rich sleep-related information including postures, hand positions, body rollover, body movements, and acoustic events. However, *SleepGuard* can not support major physiological parameters (*e.g.*, heartbeat rate and breathing rate) measurements that play an important role in people's sleep health.
iii) Ear-worn. Ear-worn devices provide new sensing opportunities for sleep monitoring [54], which can measure rich physiological information [45] including brain waves [16], heartbeat [8], blood pressure [7], and respiration [14]. For example, prior works [42, 43, 63] have conducted overnight sleep-related studies using in-ear EEG sensors. Nevertheless, such biosensors are not widely equipped on most commercial devices. Hence, some studies have started to leverage existing sensors on earbuds to capture physiological information (*e.g.*, heartbeat and breathing) that plays an important role in sleep. For example, some earable-based studies have explored the feasibility of physiological monitoring using passive acoustic modality [8, 9, 24] and active acoustic modality [62].

However, leveraging existing available sensors on earbuds to achieve fine-grained sleep stage detection is still lacking. In order to fill the gap in this aspect, we propose EarSleep that utilizes low-cost in-ear microphones on commercial earbuds to enable fine-grained sleep monitoring and analysis.

9 CONCLUSION

With society's emphasis on sleep health and the fast development of the sleep earbuds market, we propose an earbud-based sleep monitoring system, named EarSleep, which leverages a pair of ubiquitous in-ear microphones for sleep activity recognition and sleep stage detection. Different from prior works, EarSleep only takes advantage of in-ear acoustic modality to derive rich sleep information from various physical and physiological activities to represent sleep stage transitions, promoting the progress of ubiquitous sensing of earbuds to mobile health monitoring.

ACKNOWLEDGMENTS

Panlong Yang and Xiang-Yang Li are the corresponding authors. This work is supported by the National Natural Science Foundation of China with No. 62132018. We appreciate Nanjing Xiaocheng Health Science & Technology Co., Ltd. for their help and support regarding sleep medicine knowledge. In addition, we also appreciate the valuable suggestions and feedback from the anonymous reviewers.

REFERENCES

[1] Greg Atkinson and Damien Davenne. 2007. Relationships between sleep, physical activity and human health. *Physiology & behavior* 90, 2-3 (2007), 229–235.

[2] R Nisha Aurora, David A Kristo, Sabin R Bista, James A Rowley, Rochelle S Zak, Kenneth R Casey, Carin I Lamm, Sharon L Tracy, and Richard S Rosenberg. 2012. The treatment of restless legs syndrome and periodic limb movement disorder in adults—an update for 2012: practice parameters with an evidence-based systematic review and meta-analyses: an American Academy of Sleep Medicine Clinical Practice Guideline. *Sleep* 35, 8 (2012), 1039–1062.

[3] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer.. In *NDSS*, Vol. 2020. 1–18.

[4] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. 2012. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* 176 (2012), 2012.

[5] MH Bonnet and DL Arand. 1997. Heart rate variability: sleep stage, time of night, and arousal influences. *Electroencephalography and clinical neurophysiology* 102, 5 (1997), 390–396.

[6] Reza Boostani, Foroozan Karimzadeh, and Mohammad Nami. 2017. A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer methods and programs in biomedicine* 140 (2017), 77–91.

[7] Nam Bui, Nhat Pham, Hoang Truong, Phuc Nguyen, Jianliang Xiao, Robin Deterding, Thang Dinh, and Tam Vu. 2020. ebp: Frequent and comfortable blood pressure monitoring from inside human's ears. *GetMobile: Mobile Computing and Communications* 23, 4 (2020), 34–38.

[8] 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)*. 200–209.

[9] Yetong Cao, Chao Cai, Fan Li, Zhe Chen, and Jun Luo. 2023. HeartPrint: Passive Heart Sounds Authentication Exploiting In-Ear Microphones. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.

[10] Kévin Carillo, Olivier Doutres, and Franck Sgard. 2020. Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America* 147, 5 (2020), 3476–3489.

[11] Rosalind D Cartwright, Frank Diaz, and Stephen Lloyd. 1991. The effects of sleep posture and sleep stage on apnea frequency. *Sleep* 14, 4 (1991), 351–353.

[12] Liqiong Chang, Jiaqi Lu, Ju Wang, Xiaojiang Chen, Dingyi Fang, Zhanyong Tang, Petteri Nurmi, and Zheng Wang. 2018. SleepGuard: Capturing rich sleep information using smartwatch sensing data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–34.

[13] Xiangmao Chang, Cheng Peng, Guoliang Xing, Tian Hao, and Gang Zhou. 2020. ISleep: A smartphone system for unobtrusive sleep quality monitoring. *ACM Transactions on Sensor Networks (TOSN)* 16, 3 (2020), 1–32.

[14] Harry J Davies, Patrik Bachtiger, Ian Williams, Philip L Molyneaux, Nicholas S Peters, and Danilo P Mandic. 2022. Wearable in-ear PPG: Detailed respiratory variations enable classification of COPD. *IEEE Transactions on Biomedical Engineering* 69, 7 (2022), 2390–2400.
[15] Konstantin Dragomiretskiy and Dominique Zosso. 2013. Variational mode decomposition. *IEEE transactions on signal processing* 62, 3 (2013), 531–544.
[16] Andrea Ferlini, Dong Ma, Lorena Qendro, and Cecilia Mascolo. 2022. Mobile Health With Head-Worn Devices: Challenges and Opportunities. *IEEE Pervasive Computing* 21, 3 (2022), 52–60.
[17] Fiorenza Giganti, Gianluca Ficca, Sara Gori, and Piero Salzarulo. 2008. Body movements during night sleep and their relationship with sleep stages are further modified in very old subjects. *Brain research bulletin* 75, 1 (2008), 66–69.
[18] Michael A Grandner. 2017. Sleep, health, and society. *Sleep medicine clinics* 12, 1 (2017), 1–22.
[19] Weixi Gu, Longfei Shangguan, Zheng Yang, and Yunhao Liu. 2015. Sleep hunter: Towards fine grained sleep stage tracking with smartphones. *IEEE Transactions on Mobile Computing* 15, 6 (2015), 1514–1527.
[20] Unsoo Ha, Salah Assana, and Fadel Adib. 2020. Contactless seismocardiography via deep learning radars. In *Proceedings of the 26th annual international conference on mobile computing and networking*. 1–14.
[21] Lauren Hale, Wendy Troxel, and Daniel J Buysse. 2020. Sleep health: an opportunity for public health to address health equity. *Annual review of public health* 41 (2020), 81–99.
[22] Feiyu Han, Panlong Yang, Haohua Du, and Xiang-Yang Li. 2023. Accuth: Anti-Spoofing Voice Authentication via Accelerometer. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 637–650.
[23] Feiyu Han, Panlong Yang, Haohua Du, and Xiang-Yang Li. 2024. Accuth$^+$+: Accelerometer-Based Anti-Spoofing Voice Authentication on Wrist-Worn Wearables. *IEEE Transactions on Mobile Computing* 23, 5 (2024), 5571–5588.
[24] Feiyu Han, Panlong Yang, Shaojie Yan, Haohua Du, and Yuanhao Feng. 2023. BreathSign: Transparent and Continuous In-ear Authentication Using Bone-conducted Breathing Biometrics. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.
[25] Chunhua He, Jiewen Tan, Xuelei Jian, Guangxiong Zhong, Lianglun Cheng, and Juze Lin. 2022. A smart flexible vital signs and sleep monitoring belt based on MEMS triaxial accelerometer and pressure sensor. *IEEE Internet of Things Journal* 9, 15 (2022), 14126–14136.
[26] Market IQ Hub. 2023. Decoding the Sleep Headphones Market: A Deep Dive into the Latest Market Trends, Market Segmentation, and Competitive Analysis. https://www.linkedin.com/pulse/decoding-sleep-headphones-market-deep-dive-latest-trends/.
[27] Zawar Hussain, Subhash Sagar, Wei Emma Zhang, and Quan Z Sheng. 2019. A cost-effective and non-invasive system for sleep and vital signs monitoring using passive RFID tags. In *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 153–161.
[28] Ali Kiaghadi, Seyedeh Zohreh Homayounfar, Jeremy Gummeson, Trisha Andrew, and Deepak Ganesan. 2019. Phyjama: Physiological sensing via fiber-enhanced pyjamas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–29.
[29] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging acoustic dispersion for ubiquitous touch interface on solid surfaces. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 211–223.
[30] Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang, and Mark D Plumbley. 2019. Sound event detection and time–frequency segmentation from weakly labelled data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 4 (2019), 777–787.
[31] Yosuke Kurihara and Kajiro Watanabe. 2012. Sleep-stage decision algorithm by using heartbeat and body-movement signals. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 42, 6 (2012), 1450–1459.
[32] Asaf Levartovsky, Eliran Dafna, Yaniv Zigel, and Ariel Tarasiuk. 2016. Breathing and snoring sound characteristics during sleep in adults. *Journal of Clinical Sleep Medicine* 12, 3 (2016), 375–384.
[33] Chen Liu, Jie Xiong, Lin Cai, Lin Feng, Xiaojiang Chen, and Dingyi Fang. 2019. Beyond respiration: Contactless sleep sound-activity recognition using RF signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–22.
[34] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing*. 267–276.
[35] Xi Long, Jérôme Foussier, Pedro Fonseca, Reinder Haakma, and Ronald M Aarts. 2014. Analyzing respiratory effort amplitude for automated sleep stage classification. *Biomedical Signal Processing and Control* 14 (2014), 197–205.
[36] Régis Lopez, Sofiène Chenini, Lucie Barateau, Anna-Laura Rassu, Elisa Evangelista, Beatriz Abril, Julien Fanielle, Nicolas Vitello, Isabelle Jaussent, and Yves Dauvilliers. 2021. Sleep-related head jerks: toward a new movement disorder. *Sleep* 44, 2 (2021), zsaa165.
[37] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 175–187.
[38] Miguel Marino, Yi Li, Michael N Rueschman, John W Winkelman, JM Ellenbogen, Jo M Solet, Hilary Dulin, Lisa F Berkman, and Orfeu M Buxton. 2013. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* 36, 11 (2013), 1747–1755.

[39] Pejman Memar and Farhad Faradji. 2017. A novel multi-class EEG-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 1 (2017), 84–95.
[40] Fábio Mendonça, Sheikh Shanawaz Mostafa, Fernando Morgado-Dias, and Antonio G Ravelo-García. 2018. Sleep quality estimation by cardiopulmonary coupling analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 12 (2018), 2233–2239.
[41] Elmar Messner, Matthias Zöhrer, and Franz Pernkopf. 2018. Heart sound segmentation—An event detection approach using deep recurrent neural networks. *IEEE transactions on biomedical engineering* 65, 9 (2018), 1964–1974.
[42] Kaare B Mikkelsen, Yousef R Tabar, Simon L Kappel, Christian B Christensen, Hans O Toft, Martin C Hemmsen, Mike L Rank, Marit Otto, and Preben Kidmose. 2019. Accurate whole-night sleep monitoring with dry-contact ear-EEG. *Scientific reports* 9, 1 (2019), 16824.
[43] Takashi Nakamura, Yousef D Alqurashi, Mary J Morrell, and Danilo P Mandic. 2019. Hearables: automatic overnight sleep monitoring with standardized in-ear EEG sensor. *IEEE Transactions on Biomedical Engineering* 67, 1 (2019), 203–212.
[44] Ltd Nanjing Xiaocheng Health Science & Technology Co. 2023. Meet Sleep. https://meetsleeping.com/.
[45] Colver Ken Howe Ne, Jameel Muzaffar, Aakash Amlani, and Manohar Bance. 2021. Hearables, in-ear sensing devices for bio-signal acquisition: a narrative review. *Expert Review of Medical Devices* 18, sup1 (2021), 95–128.
[46] ALISTER McKENZIE Neill, Susan Michelle Angus, Dimitar Sajkov, and RONALD DOUGLAS McEVOY. 1997. Effects of sleep posture on upper airway stability in patients with obstructive sleep apnea. *American journal of respiratory and critical care medicine* 155, 1 (1997), 199–204.
[47] Aakash K Patel, Vamsi Reddy, and John F Araujo. 2022. Physiology, sleep stages. In *StatPearls [Internet]*. StatPearls Publishing.
[48] Thomas Penzel, Jan W Kantelhardt, Chung-Chang Lo, Karlheinz Voigt, and Claus Vogelmeier. 2003. Dynamics of heart rate and sleep stages in normals and patients with sleep apnea. *Neuropsychopharmacology* 28, 1 (2003), S48–S53.
[49] Ignacio Perez-Pozuelo, Marius Posa, Dimitris Spathis, Kate Westgate, Nicholas Wareham, Cecilia Mascolo, Søren Brage, and Joao Palotti. 2022. Detecting sleep outside the clinic using wearable heart rate devices. *Scientific Reports* 12, 1 (2022), 7956.
[50] Nuno Pombo and Nuno M Garcia. 2016. ubiSleep: An ubiquitous sensor system for sleep monitoring. In *2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 1–4.
[51] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.
[52] Brandt Ranj and Sage Anderson. 2023. Best Earbuds For Sleep 2023: Reviews of Top-Rated Sleep Earbuds Online. https://www.rollingstone.com/product-recommendations/electronics/best-sleep-earbuds-1146657/.
[53] Yanzhi Ren, Chen Wang, Jie Yang, and Yingying Chen. 2015. Fine-grained sleep monitoring: Hearing your breathing with smartphones. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1194–1202.
[54] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with earables: A systematic literature review and taxonomy of phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–57.
[55] Tobias Röddiger, Christian Dinse, and Michael Beigl. 2021. Wearability and comfort of earables during sleep. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 150–152.
[56] Avi Sadeh. 2011. The role and validity of actigraphy in sleep medicine: an update. *Sleep medicine reviews* 15, 4 (2011), 259–267.
[57] Malay Sarkar, Irappa Madabhavi, Narasimhalu Niranjan, and Megha Dogra. 2015. Auscultation of the respiratory system. *Annals of thoracic medicine* 10, 3 (2015), 158.
[58] Abraham Savitzky and Marcel JE Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36, 8 (1964), 1627–1639.
[59] Aicko Y Schumann, Ronny P Bartsch, Thomas Penzel, Plamen Ch Ivanov, and Jan W Kantelhardt. 2010. Aging effects on cardiac and respiratory dynamics in healthy subjects across sleep stages. *Sleep* 33, 7 (2010), 943–955.
[60] Pragya Sharma and Edwin C Kan. 2018. Sleep scoring with a UHF RFID tag by near field coherent sensing. In *2018 IEEE/MTT-S International Microwave Symposium-IMS*. IEEE, 1419–1422.
[61] Xiao Sun, Li Qiu, Yibo Wu, Yeming Tang, and Guohong Cao. 2017. Sleepmonitor: Monitoring respiratory rate and body position during sleep using smartwatch. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–22.
[62] Xue Sun, Jie Xiong, Chao Feng, Wenwen Deng, Xudong Wei, Dingyi Fang, and Xiaojiang Chen. 2023. Earmonitor: In-ear Motion-resilient Acoustic Sensing Using Commodity Earphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–22.
[63] Yousef Rezaei Tabar, Kaare B Mikkelsen, Mike Lind Rank, Martin Christian Hemmsen, Marit Otto, and Preben Kidmose. 2021. Ear-EEG for sleep assessment: a comparison with actigraphy and PSG. *Sleep and Breathing* 25 (2021), 1693–1705.
[64] Xudong Tan, Menghan Hu, Guangtao Zhai, Yan Zhu, Wenfang Li, and Xiao-Ping Zhang. 2023. Lightweight Video-Based Respiration Rate Detection Algorithm: An Application Case on Intensive Care. *IEEE Transactions on Multimedia* (2023).
[65] Robert Joseph Thomas, Joseph E Mietus, Chung-Kang Peng, and Ary L Goldberger. 2005. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. *Sleep* 28, 9 (2005), 1151–1161.

[66] EA Tom Taulli. 2022. How Much Does the Average Sleep Study Cost? - GoodRx. https://www.goodrx.com/health-topic/procedures/how-much-sleep-study-cost.
[67] Juergen Tonndorf. 1968. A new concept of bone conduction. *Archives of Otolaryngology* 87, 6 (1968), 595–600.
[68] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[69] Sudip Vhaduri. 2020. Nocturnal cough and snore detection using smartphones in presence of multiple background-noises. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 174–186.
[70] Yunlu Wang, Cheng Yang, Menghan Hu, Jian Zhang, Qingli Li, Guangtao Zhai, and Xiao-Ping Zhang. 2021. Identification of deep breath while moving forward based on multiple body regions and graph signal analysis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7958–7962.
[71] Lutz Welling and Hermann Ney. 1998. Formant estimation for speech recognition. *IEEE Transactions on Speech and Audio Processing* 6, 1 (1998), 36–48.
[72] Wikipedia. 2023. Detrended fluctuation analysis. https://en.wikipedia.org/wiki/Detrended_fluctuation_analysis.
[73] Johanna Wilde-Frenz and Hartmut Schulz. 1983. Rate and distribution of body movements during sleep in humans. *Perceptual and motor skills* 56, 1 (1983), 275–283.
[74] Yanni Yang, Jiannong Cao, Xuefeng Liu, and Kai Xing. 2018. Multi-person sleeping respiration monitoring with COTS WiFi devices. In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 37–45.
[75] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2017. Vital sign and sleep monitoring using millimeter wave. *ACM Transactions on Sensor Networks (TOSN)* 13, 2 (2017), 1–32.
[76] Bohan Yu, Yuxiang Wang, Kai Niu, Youwei Zeng, Tao Gu, Leye Wang, Cuntai Guan, and Daqing Zhang. 2021. WiFi-Sleep: sleep stage monitoring using commodity Wi-Fi devices. *IEEE internet of things journal* 8, 18 (2021), 13900–13913.
[77] Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. 2018. Extracting multi-person respiration from entangled RF signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–22.
[78] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. 2020. BodyCompass: Monitoring sleep posture with wireless signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–25.
[79] Youwei Zeng, Dan Wu, Jie Xiong, Jinyi Liu, Zhaopeng Liu, and Daqing Zhang. 2020. MultiSense: Enabling multi-person respiration sensing with commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.
[80] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.
[81] Feng Zhang, Chenshu Wu, Beibei Wang, Min Wu, Daniel Bugos, Hangfang Zhang, and KJ Ray Liu. 2019. SMARS: Sleep monitoring via ambient radio signals. *IEEE Transactions on Mobile Computing* 20, 1 (2019), 217–231.
[82] Youwei Zhang, Feiyu Han, Panlong Yang, Yuanhao Feng, Yubo Yan, and Ran Guan. 2023. Wi-Cyclops: Room-Scale WiFi Sensing System for Respiration Detection Based on Single-Antenna. *ACM Transactions on Sensor Networks* (2023).