WakeUp: Fine-Grained Fatigue Detection Based on Multi-Information Fusion on Smart Speakers

Zhiyuan Zhao* Fan Li* Yadong Xie* Yu Wang†
* School of Computer Science, Beijing Institute of Technology, Beijing, China.
† Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA.
Email: {3120205519, fli, ydxie}@bit.edu.cn, wangyu@temple.edu

***Abstract*—With the development of society and the gradual increase of life pressure, the number of people engaged in mental work and working hours have increased significantly, resulting in more and more people in a state of fatigue. It not only reduces people's work efficiency, but also causes health and safety related problems. The existing fatigue detection systems either have different shortcomings in diverse scenarios or are limited by proprietary equipment, which is difficult to be applied in real life. Motivated by this, we propose a multi-information fatigue detection system named WakeUp based on commercial smart speakers, which is the first to fuse physiological and behavioral information for fine-grained fatigue detection in a non-contact manner. We carefully design a method to simultaneously extract users' physiological and behavioral information based on the MobileViT network and VMD decomposition algorithm respectively. Then, we design a multi-information fusion method based on the statistical features of these two kinds of information. In addition, we adopt an SVM classifier to achieve fine-grained fatigue level. Extensive experiments with 20 volunteers show that WakeUp can detect fatigue with an accuracy of 97.28%. Meanwhile, WakeUp can maintain stability and robustness under different experimental settings.**

I. INTRODUCTION

Fatigue is a physiological phenomenon caused by excessive physical or mental work [1]. It reduces people's work efficiency, has a serious impact on physical and mental health, and causes safety-related problems, resulting in serious economic losses. In the past two years, the proportion of employees who often work remotely has soared from 23% before the outbreak of COVID-19 to 71% at present [2]. According to the surveys [3], [4], about 69% of remote workers suffer from fatigue, which greatly reduces work efficiency. In addition, fatigue can lead to obesity, heart disease, depression, some cancers, sleep disorders, and other problems. It also reduces employees' immunity to viruses and increases the possibility of cold or flu transmission. According to statistics, the health problems caused by fatigue cost American companies about 136.4 billion dollars a year [5]. Therefore, how to effectively and accurately detect fatigue and timely warning are of great practical significance.

In recent decades, researchers constantly try to find an effective method to detect fatigue. One method to detect fatigue is called subjective fatigue detection, which can identify whether people are fatigued through medical observation and research, or inspection judgment, response test, and other means. Common subjective fatigue detection methods include Fatigue Assessment Scale (FAS) [6], Karolinska Sleep Scale (KSS) [7], and so on. In addition, another method to detect fatigue is based on objective symptoms. When people are fatigued, they can show obvious symptoms of fatigue, which are usually reflected in people's physiology and behavior, such as slow respiration, nodding, yawning, and other symptoms [8]. Judging people's fatigue levels by detecting fatigue-related symptoms has become a hot research direction concerned by many research institutions.

On one hand, the fatigue level can be determined by detecting the changes of people's physiological indicators (e.g., EEG, ECG, EMG) [9], [10]. Physiological signals can truly reflect people's fatigue state and have the advantage of high reliability. However, traditional methods need to attach multiple sensing devices on the human body, so the measurement is inconvenient and the practicability is limited. On the other hand, vision-based detection methods evaluate the fatigue level according to the behavioral information of people's head and face (e.g., eyes, mouth) in the fatigue state [11], [12], which has the advantage of non-contact detection. Nevertheless, due to the acquisition of facial image information, there is a problem of user privacy leakage. Therefore, a non-contact, low-cost, easy-to-deploy, and fine-grained fatigue detection system is urgently needed to improve people's health levels and work efficiency.

To this end, we further investigate the feasibility of using acoustic sensing for fatigue detection, which has been widely used in indoor location [13], [14] and human computer interface [15], [16]. In addition, smart speakers that can record our daily commands, such as Amazon Echo, Google Home, and Apple HomePod, have become common in homes around the world. By the end of 2021, the total number of smart speakers in the world has reached nearly 163 million. According to Canalys' latest forecast, the global installation base of smart speakers is expected to reach 640 million by 2024 [17]. The widespread adoption of high-quality smart speakers with multiple microphones also provides a unique opportunity for non-contact fatigue detection. In this paper, we design and implement WakeUp, the first multi-information fusion fatigue detection system at home/office environment based on

Fan Li is the corresponding author. The work of Fan Li is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62072040.

commercial smart speakers. WakeUp not only monitors the physiological information (i.e., respiration) of fatigue, but also integrates the behavioral information (i.e., yawning, nodding, stretching, and sighing) of fatigue.

The key idea of WakeUp is to transmit inaudible high-frequency ultrasonic signals through smart speakers, which are reflected by users and received by the microphone array, and then process the transmitted and received signals to extract the fatigue statistics of users. To realize WakeUp, several challenges need to be addressed. Firstly, at home/office, the signals reflected by users are easily disturbed by the activities (such as sweeping, walking, etc.) of people around them. In order to minimize the impact of these interferences, we take advantage of the property that smart speakers have multiple microphones to locate users. Since the distance between the microphones is very close, triangulation may be greatly limited. In addition to distance, another key parameter related to user's location is the direction of arrival (DOA) of the signal. We use generalized cross-correlation with phase transform (GCC-PHAT) [18] to obtain the accurate direction of the user and further use time-delay beamforming to amplify the reflected signal.

Secondly, when we get the user's reflected signal, how to process it to extract the user's physiological and behavioral information? We construct a virtual transmitted signal and iteratively mix it with the reflected signal to eliminate the random time delay from the signal is triggered to be sent until the signal is actually sent out and the confusion between the direct path and the target reflection path [19]. Then the mixed signal is low-pass filtered to obtain the intermediate frequency (IF) signal [20]. Next, we first segment the IF signal, then perform time-reassigned multisynchrosqueezing transform (TMSST) [21] and wavelet scattering [22], [23] for feature extraction, and finally send the wavelet scattering results to a designed MobileViT [24] to obtain user's behavioral information. We simultaneously extract the phase information of the IF signal and use band-pass filtering to remove the influence of behavioral information and high-frequency noises. Then we further use the variational mode decomposition (VMD) [25] to obtain user's physiological information. After obtaining the user's behavioral and physiological information, the last challenge is how to fuse this information and get the user's fine-grained fatigue detection results. To this end, we extract the statistics of this information and use an SVM classifier to get the current fine-grained fatigue state of users.

We implement WakeUp using a Raspberry Pi 4B, a regular hexagonal 6-microphone array, an omni-directional speaker, a respiration belt, and a laptop. We recruit 20 volunteers (11 males and 9 females) and conduct data collection in four different environments for evaluation. We end up collecting 2000 samples containing fatigue information to evaluate WakeUp, and each sample contains a 2-minute signal. Results demonstrate that WakeUp can accurately detect fatigue under different experimental settings.

Our contributions are summarized as follows:

- To the best of our knowledge, WakeUp is the first system that uses smart speakers to fuse physiological and behavioral information to achieve fine-grained fatigue detection. Compared to the single-information solutions, our system achieves better performance, robustness, and universality.
- We apply the features of microphone array to make WakeUp have a good anti-interference ability. We carefully design a feature extraction scheme based on TMSST, wavelet scattering and the lightweight network MobileViT to obtain the user's behavioral information. Moreover, we also design an extraction scheme based on VMD to obtain physiological information.
- Based on statistical information, we design a fusion scheme of behavioral and physiological information and further combine with an SVM classifier to achieve fine-grained fatigue detection.
- We implement a prototype of WakeUp and conduct extensive experiments to evaluate its effectiveness with 20 volunteers. The results show that WakeUp can detect fatigue with an average accuracy of 97.28%. Moreover, we verify the stability of WakeUp under different experimental settings, and the results suggest the system can maintain the superiority of the performance.

II. RELATED WORK

Recent works in fatigue detection can be divided into contact-based method, non-contact-based method, and hybrid method.

Contact-based method. Contact methods mainly obtain behavioral and physiological information through proprietary equipment. Yeo et al. [26] use the "double banana" bipolar referencing montage with 19 channels to collect EEG signals and conduct nonlinear analysis on them, so as to determine the user's fatigue level. Katsis et al. [27] employ the newly developed seat to record the user's surface EMG signal and then reflect the user's fatigue state based on the amplitude and frequency characteristics of the EMG signal. NodeTrack [28] exploits the phase difference between two RFID tags mounted on the back of the hat worn by the user to extract fatigue-related nodding features. However, these works require additional deployment of costly and environmentally sensitive equipment, and contact-based signal acquisition also brings comfort issues.

Non-contact-based method. Non-contact solutions mainly obtain physiological and behavioral information through visual and wireless signals. Geng et al. [29] utilize infrared acquisition equipment to collect user facial images, combine with the cascade regression method to locate the feature points of eyes and mouth, and use a neural network for fatigue detection. However, vision-based methods may lead to privacy leakage, and the detection accuracy is easily affected by light and obstacles. In addition, Liu et al. [30] obtain respiration and heartbeat through millimeter wave radar and built a back-propagation neural network model to predict fatigue. WiFind [31] studies the impact of physical features on WiFi signals when users are fatigued, and verifies the feasibility of detecting

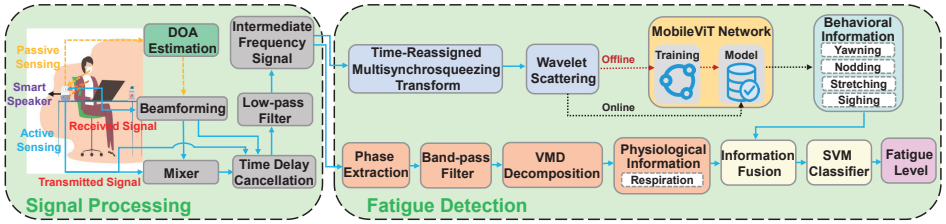


Fig. 1: System architecture of WakeUp.

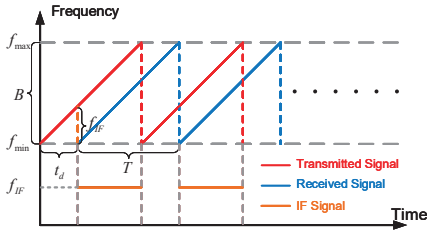


Fig. 2: FMCW signal.

fatigue through WiFi signals. However, these wireless-signal-based methods are usually very sensitive to the interference of the surrounding environment, and their anti-interference ability is relatively weak.

Hybrid method. There are also some researches focusing on using both contact and non-contact methods to extract multiple information to detect fatigue. Amirudin et al. [32] exploit the physiological features of theta band in EEG signals and combine with the eye state to detect fatigue. Abbas et al. [33] develop a new hybrid fatigue detection system based on multiple cameras and ECG sensors combined with visual and non-visual features. Craye et al. [34] use multiple sensors to extract multiple features (e.g., sound, image, heart rate) to judge the fatigue state of users. But these works generally require a variety of sensing equipment, which results in poor system deployment capability. However, compared with single information, the multiple information can contain more features.

Unlike the above works, WakeUp automatically obtains multiple information of users through acoustic sensing to detect fatigue without wearing extra equipment. WakeUp has good anti-interference ability by taking advantage of omni-directional speaker and microphone array.

III. SYSTEM DESIGN

This section details the design of WakeUp. We also highlight the key observations and core techniques behind the detection of fatigue.

A. System Overview

Fig. 1 shows the architecture of WakeUp. The whole system includes 2 parts, *signal processing* and *fatigue detection*.

In *Signal Processing*, the user first sends the system startup command, then the microphone array records the command signal and the system starts to work. In order to eliminate the interference of the activities of surrounding people, we process the command signals recorded by these microphones to obtain the DOA of the signals. Next, the speaker emits frequency modulated continuous wave (FMCW) signals, and then we use beamforming to synthesize multiple received signals into one signal to increase the strength of the reflected signal and suppress noise. After that, we mix this synthetic signal with the transmitted signal. Furthermore, we perform time delay cancellation on the mixed signal to eliminate the time delay caused by the hardware system from the signal is triggered to be sent until the signal is actually sent out and the confusion between the direct path and the target reflection path. Finally, we utilize a low-pass filter on this signal to get the IF signal reflected by the user.

In *Fatigue Detection*, we analyze the IF signal to obtain behavioral information and physiological information of fatigue. In order to obtain behavioral information, we first perform TMSST on the IF signal, and then we perform wavelet scattering transform on the transformed results to extract translation-invariant, stable, and informative features. Next, we feed these features into a MobileViT lightweight neural network to get four kinds of behavioral information, including yawning, nodding, stretching, and sighing. In addition, in order to obtain physiological information, we first extract the phase information of the IF signal and further obtain the phase change. Then we perform band-pass filtering on the phase change signal to remove the influence of low-frequency behavioral information and high-frequency noises. Next, we perform VMD decomposition on the filtered phase change and Fast Fourier Transform (FFT) on each component to obtain physiological information (i.e., respiration). After getting the user's behavioral and physiological information, we calculate their statistical features for a period of time. Finally, we employ these statistical features to train an SVM classifier to achieve fine-grained fatigue detection.

B. Signal Processing

DOA Estimation: When users work at home/office, the signals reflected by them are easily disturbed by the activities of people around them. To eliminate these interferences, we use the microphone array to locate the user. Since the user hardly changes the location during most working hours, we only need to calculate the user's direction relative to the smart speaker when the system starts to work. Firstly, the user sends the startup command to the smart speaker, then the microphone array records this command signal and the system starts to work. For convenience, we can establish a three-dimensional Cartesian coordinate system with the center of the smart speaker as the origin. The direction of the user relative to the smart speaker can be determined with two parameters, namely, azimuth angle and elevation angle. The azimuth angle is defined as the angle, in the xy-plane, from the x-axis toward the y-axis. The elevation angle is defined as the angle from the xy-plane toward the z-axis. Since the command signal arrives at each microphone at different time, we estimate the direction of the user based on Time Difference of Arrival (TDOA). We use the GCC-PHAT algorithm to calculate the time delay, which assumes that the signal source is located in the far field of the array, so the DOA is the same for

all microphones. We use the command signals from the user received by the microphone array, and estimate the correlation between each signal pair via GCC-PHAT. The maximum peak in each correlation is further found to determine the delay between these two signals. Finally, we estimate the azimuth and elevation angles using least-square estimation.

FMCW Signal Design: After obtaining the DOA, the smart speaker starts to transmit FMCW signals, which are reflected by the user and received by the microphone array. The basic principle of FMCW is to transmit frequency continuous wave and its frequency changes with time as shown in Fig. 2. The transmitted signal x_T with frequency conversion period T can be expressed as

$$x_T(t) = A_T \cos\left(2\pi\left(f_{\min} + \frac{B}{2T}t\right)t\right), t \in (0, T), \quad (1)$$

where A_T represents the amplitude of the transmitted signal. $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequency of FMCW respectively. The sound frequency above $15kHz$ is already inaudible for most adults [35], and considering the upper limit of sound frequency produced by most smart speakers, we set $f_{\min}$ and $f_{\max}$ to $16kHz$ and $21kHz$, respectively. $B = f_{\max} - f_{\min}$ is the bandwidth of the FMCW. When the speaker and microphone array are co-located and the distance between the target and the sensing device is R, the signal reaches the target, and then reflects back from the target to the microphone array after time period $t_d = 2R/c$, where c is the propagation speed of the signal in the air. Therefore, the signal $x_R(t)$ received by the microphone array can be expressed as

$$x_R(t) = A_R \cos\left(2\pi\left(f_{\min} + \frac{B}{2T}(t - t_d)\right)(t - t_d)\right), \quad (2)$$

where A_R is the amplitude of the received signal. Finally, we multiply $x_T(t)$ and $x_R(t)$ to mix, and pass the mixed signal through the low-pass filter to obtain the IF signal x_{IF} containing the target information, whose frequency is f_{IF} [20].

Beamforming: Now we have not only the azimuth and elevation angles of the user relative to the smart speaker, but also the reflected signal received by the microphone. Next, we use time-delay beamforming to suppress the interference of surrounding people and enhance the strength of the user's reflected signal. The time delay beamformer performs delay-and-sum beamforming, which can compensate a reflected signal coming from a specific direction for the arrival time differences across the microphones. Reflected signals arriving at the array elements are time-aligned and then summed. Time alignment is achieved by transforming the signals into the frequency domain and applying linear phase shifts corresponding to a time delay. The individual signals are then added and converted back to the time domain. Finally, the signals received by the six microphones are combined into one signal.

Time Delay Cancellation: There is a random time delay from the signal is triggered to be sent until the signal is actually sent out and the sequence of the direct path signal and the target reflection path signal in time can be chaotic. The direct path means that the signal directly reaches the microphone array from the speaker, while the target reflection path means that the signal is emitted by the speaker, then reflected by the target, and finally received by the microphone array. In order to eliminate this delay and path confusion, we adopt a method of iteratively mixing the virtual transmitted signal and the received signal. We first perform FFT on the filtered signal, and then select the maximum peak f_d to obtain the delay $t_v = f_d T/2B$ to construct the virtual transmission signal x_V:

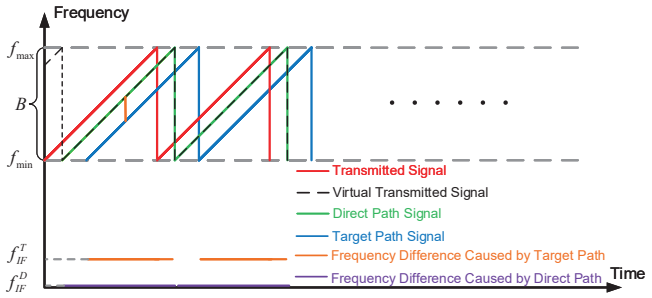


Fig. 3: Time delay cancellation.

$$x_V(t) = A_T \cos\left(2\pi\left(f_{\min} + \frac{B}{2T}(t + t_v)\right)(t + t_v)\right). \quad (3)$$

Next, x_V and x'_R are mixed, and low-pass filtering and FFT are carried out again. x'_R represents the signal corresponding to x_V received by the microphone array and beamformed. If the maximum peak is located at the timestamp "0", it means that the delay is successfully eliminated, otherwise, let $t_v = T - t_v$, we continue to construct a virtual transmission signal for iteration until the maximum peak is located at the timestamp "0" [19]. Generally, the above process needs to be iterated twice at most, and the schematic diagram of the results after iteration is shown in Fig. 3. It can also be seen from the figure that this process also solves the problem of confusion between the direct path and the reflection path. Finally, we mix x_V and x'_R and perform low-pass filtering to obtain IF signal x_{IF}^T containing user's fatigue information, which can be expressed as:

$$x_{IF}^U = \frac{A_T A_R}{2} \cos\left(2\pi\frac{B}{T}t t_d + 2\pi f_{\min} t_d - \pi\frac{B}{T}t_d^2\right). \quad (4)$$

From the above equation, it can be concluded that the phase of the signal is $\varphi = -2\pi\left(f_{\min} t_d - \frac{B}{2T}t_d^2\right)$. Usually, $f_{\min} \gg Bt_d/2T$, so we can omit the quadratic term. By replacing t_d with $2R/c$, the phase change $\Delta\varphi$ caused by the change of path length ΔR can be expressed as $\Delta\varphi = -\frac{4\pi f_{\min}\Delta R}{c}$.

C. Fatigue Detection

After obtaining the IF signal containing user's fatigue information, we simultaneously process the IF signal and its phase change to get behavioral and physiological information respectively. Since training a network and an SVM classifier requires a large amount of data, it is usually necessary to collect enough data before information extraction. Therefore, we first introduce the process of data collecting, then describe

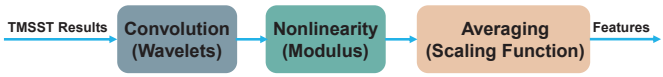


Fig. 4: Wavelet scattering framework.

the specific methods of two kinds of information extraction in detail, and finally give the algorithm of information fusion and fatigue detection.

1) Data Collecting: We recruit 20 volunteers, including 11 males and 9 females. We collect data with smart speakers in office, dormitory, bedroom, and study room for two months. After the above *Signal Processing* process, we obtain about 2000 samples containing fatigue information, each containing a 2-minute signal. In addition, we use a camera to record the user's real behavioral information and employ a respiration belt to get the groundtruth of the user's respiration.

2) Behavioral Information Extraction: After analyzing these sample data, it is found that about 96% of yawning, nodding, sighing, and stretching can be completed in $2.9s$, $2.4s$, $2.8s$, and $5s$ respectively. Before feature extraction, we divide these 2-minute signals into frames, and the frame length is equal to $0.2s$.

Feature Extraction: We first perform TMSST on the IF signal x_{IF}^{U} containing user's fatigue information obtained in III-B. TMSST is a time-frequency analysis method, which owns the capability to provide an energy concentrated time-frequency representation result for the frequency-varying signal. We use short-time Fourier transform (STFT) to extend x_{IF}^{U} to the time-frequency domain. In the frequency domain, the STFT using the moving window function $\hat{g}(\xi)$ signal can be written as:

$$G(u,t) = (2\pi)^{-1} \int_{-\infty}^{+\infty} x_{IF}^{U}(\xi)\, \hat{g}(\xi - t)\, e^{i(\xi-t)u} d\xi. \quad (5)$$

Next, $G(u,t)$ is integrated in one dimension along the time direction to compress fuzzy time-frequency energy into group delay (GD) trajectory. This process can be expressed as:

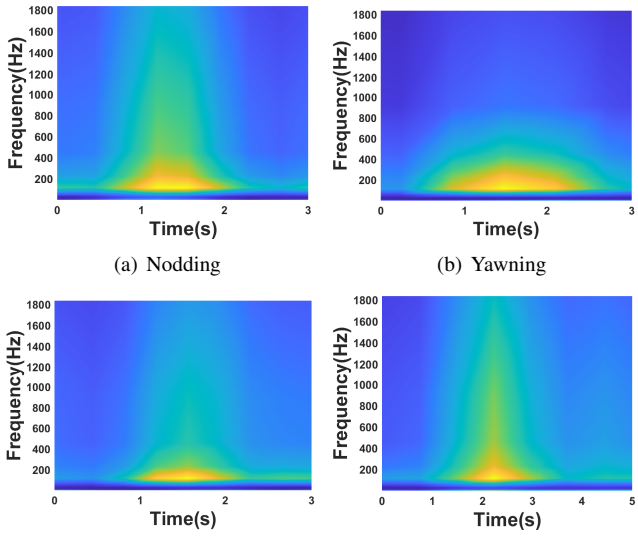
$$Tm(a,t) = \int_{-\infty}^{+\infty} G(u,t)\, \delta\left(a - \hat{t}(u,t)\right) du. \quad (6)$$

where $\delta(\cdot)$ is the Dirac delta function, and $\hat{t}(u,t)$ is the $2D$ GD estimated in [36]. In TMSST, the $2D$ GD is obtained by iteration, assuming that $\hat{t}^{N}(u,t)$ is the $2D$ group delay obtained after N iterations. We use $\hat{t}^{N}(u,t)$ to replace $\hat{t}(u,t)$ in Eq. 6:

$$Tm^{N}(a,t) = \int_{-\infty}^{+\infty} G(u,t) \cdot \delta\left(a - \hat{t}^{[N]}(u,t)\right) du. \quad (7)$$

If Eq. 7 is iterated for a sufficient number of times, the TMSST result of the IF signal can be obtained. By default, we set the number of iterations N to 150.

TMSST can provide energy concentrated time-frequency representation results for frequency change signals, but the extracted features are not stable to obtain behavioral information. Therefore, we use wavelet scattering transform to make up for this deficiency. The effect of applying wavelet scattering transform directly to the frequency-varying signal is not obvious, but we find that applying it to the result of TMSST can not only obtain stable features, but also make



(a) Nodding (b) Yawning

(c) Sighing (d) Stretching

Fig. 5: Illustration of feature extraction.

the extracted features more prominent and have the property of translation invariance. Wavelet scattering is a technology that can be used to automatically extract low variance and compact features, which can minimize the differences within classes while preserving the distinguishability between classes. As shown in Fig. 4, the wavelet scattering framework is mainly composed of three parts, including convolution, nonlinearity, and averaging. The local translation invariant feature of the input signal can be obtained by convolution, and the high-frequency information is lost in this step, which can be recovered by subsequent wavelet modulus processing. Then a scaling filter is used to average each mode to obtain the first-order scattering coefficient. By repeating the above process, we can obtain a feature matrix, which covers all orders of scattering coefficients to describe the features of the input signal. We design a wavelet time scattering transform with two filter banks. The first filter bank has a quality factor of eight wavelets per octave. The second filter bank has a quality factor of one wavelet per octave.

For each frame, we first calculate its TMSST and then perform wavelet scattering transform on the result. Fig. 5 shows the results of feature extraction of the behavioral information, and it can be seen that their frequency domain ranges are different. And obviously, we find that the frequency is mainly concentrated in the time of $1 \sim 2.5s$. Therefore, in order to obtain features of behavioral information, we use 15 frames, which can focus on the latest $3s$ signals to extract the features of yawning, nodding, sighing, and stretching. In addition, we use the t-SNE [37] algorithm to better visualize these features, as shown in Fig. 6, which shows that these behaviors can be well separated, and also reflects the effectiveness of the designed feature extraction algorithm.

Behavioral Information Acquisition: After feature extraction, we use a neural network to get this behavioral information. Our proposed network architecture is based on MobileViT, which is a light-weight, general-purpose, and

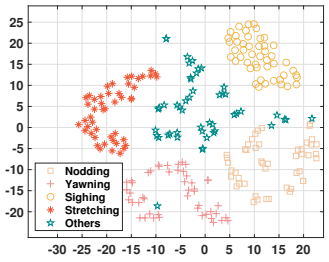


Fig. 6: Visualization of extracted features.

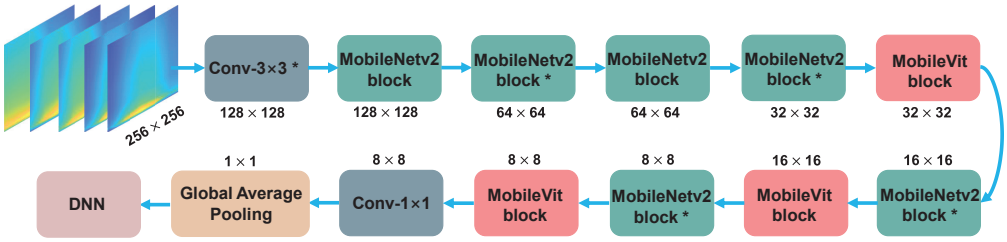


Fig. 7: Architecture of MobileViT. $*$ indicates down-sampling.

mobile-friendly transformer-based network. It combines the local representation of CNN-based networks in learning space and vision transformation (ViT) in learning global features. The architecture of MobileViT is shown in Fig. 7, which contains two important components, MobileViT block [24] and MobileNetV2 block [38]. MobileViT block is composed of three sub-modules, namely local information coding module, global information coding module, and feature fusion module. Their main structures are normal convolution and visual transformer. The corresponding functions of the three sub-modules are to extract local feature, global feature, and fusion feature. MobileViT block can fully extract features of images with fewer parameters. The main structure of MobileNetV2 block is normal convolution and depth-wise convolution. MobileNetV2 block first increases and then decreases the number of channels of feature map, and connects the residual module afterwards, which can obtain higher detection accuracy with less computation. At the end of the network, we map the output to four kinds of behavioral information through a DNN layer.

The input image size of the proposed model is 256×256. In order to unify the input, we preprocess the feature image and scale it to the input size of the model. We use the features extracted from 15 frame signals mentioned above to construct a dataset, of which 60% are used to build a train set and the remaining 40% are used to build a test set. In the online detection phase, we first take the current frame and previous 14 frames as input to extract nodding, yawning, sighing, and stretching. Then the DNN layer maps the output to a probability vector, and finally we choose a behavior with the greatest probability as the behavioral information.

3) Physiological Information Extraction: The phase change $\Delta\varphi$ includes not only the changes caused by chest displacement, but also the greater phase changes caused by users' behavioral information, so that the phase changes caused by chest displacement are submerged and cannot be extracted. Fortunately, through our study of users' behavioral information, we find that the occurrence frequency range of these behaviors is generally between $0 \sim 0.08Hz$, while the respiration frequency of normal people is generally between $0.2 \sim 0.4Hz$. Therefore, we first perform band-pass filtering on the phase change to remove the influence of behavioral information and high-frequency noises. Secondly, the frequency range of users' slow movement, such as leaning forward and backward, is generally between $0.1 \sim 2Hz$, and we cannot remove it by filtering, so we adopt VMD to decompose the filtered phase change $\Delta\varphi'$ to obtain the respiration signal. Assuming that the number of intrinsic mode function (IMF) obtained by decomposition is K, the variational constraint

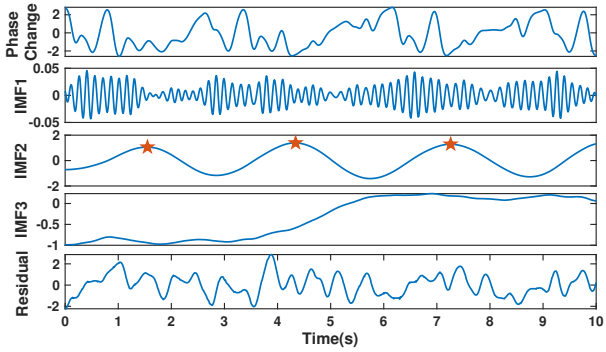


Fig. 8: Physiological information extraction.

model can be expressed as:

$$\min_{\{\zeta_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\upsilon(t) + \frac{j}{\pi t} \right) * \zeta_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \quad s.t. \sum_{k=1}^{K} \zeta_k(t) = \Delta\varphi' \tag{8}$$

where ∂_t is the partial derivative with respect to time t, $\upsilon(t)$ is the unit pulse function, j is the imaginary unit, $*$ is the convolution operation, $\zeta_k(t)$ is the k-th IMF, and $\omega_k(t)$ is the center frequency of $\zeta_k(t)$. The second-order penalty factor β and Lagrange multiplication operator $\lambda(t)$ are introduced to transform the constrained variational problem into an unconstrained problem. The augmented Lagrange function is expressed as follows:

$$\begin{aligned} L(\{\zeta_k\},\{\omega_k\},\lambda) = \beta \sum_{k=1}^{K} \left\| \partial_t \left[\left(\upsilon(t) + \frac{j}{\pi t} \right) * \zeta_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ + \left\| \Delta\varphi - \sum_{k=1}^{K} \zeta_k(t) \right\|_2^2 + \left\langle \lambda(t), \Delta\varphi' - \sum_{k=1}^{K} \zeta_k(t) \right\rangle \end{aligned} \tag{9}$$

The saddle point of the augmented Lagrange function can be obtained by iterating with alternating direction method of multipliers. K is set to 3 by default. Based on the above VMD decomposition method, we decompose the filtered phase change signal $\Delta\varphi'$ into 3 IMFs and a residual signal, and the decomposition results are shown in Fig. 8. Through the FFT of these three IMF components, it is found that IMF1 is a high-frequency component, IMF2 is between $0.21 \sim 0.45Hz$, and IMF3 is a low-frequency component. Therefore, IMF2 mainly contains respiration information. IMF1, IMF3, and residual signals mainly include the noises caused by the user's own movement. In this way, we get the physiological information of users. Moreover, we use the maximum peak search method to obtain the number of respiration, which are marked with pentagrams in Fig. 8.

4) Information Fusion and Fatigue Detection: After obtaining behavioral and physiological information, we first extract

TABLE I: Fatigue assessment scale.

Symptom	Fatigue Level
Wide awake	L_1
Relatively awake	L_2
A little fatigue	L_3
Fatigue but struggling to stay awake	L_4
Extreme fatigue and difficulty to stay awake	L_5

the statistical features of them and then use an SVM classifier for fine-grained fatigue detection.

Respiration has been used to detect fatigue [39], and recent studies find that respiration rate variability (RRV) [40] is also closely related to fatigue, so we extract some statistical features of RRV to detect fatigue. RRV refers to small changes between respiration intervals. Usually, it takes more than two minutes to measure the variability features to get a more accurate analysis, which is also the reason why each of data samples is a 2-minute signal. We have obtained the position of the maximum peak of the respiration signal in III-C3, and the interval between adjacent peaks is expressed by RR. We utilize RR to calculate five features of RRV, including $Mean\_RR$, $SDNN$, $RMSSD$, SD_1, and SD_2. $Mean\_RR$ is the mean of RR intervals, $SDNN$ is the standard deviation of RR intervals, $RMSSD$ is the square root of the mean of the sum of difference of successive RR intervals, SD_1 is the Poincaré plot standard deviation perpendicular the line of identity, and SD_2 is the Poincaré plot standard deviation along the line of identity.

After obtaining the statistical features of physiological information, we next extract the statistical features of behavioral information. We mainly count the occurrence times of these four behaviors within 2 minutes as statistical features. Through our observation, we find that the time interval between adjacent fatigue behaviors is much longer than the duration of these behaviors. Therefore, we start from detecting a certain behavior in the 15 frame signal and end when the behavior is not detected in the 15 frame signal and the whole process represents that this behavior has occurred once. In this way, we can get the times of these behaviors in 2 minutes, which are denoted as $T_{Nodding}$, $T_{Yawning}$, $T_{Sighing}$ and $T_{Stretching}$. Then we use an SVM classifier to detect fatigue, and use (X_i, L_i) to represent the training sample, where $X_i = \left[\begin{array}{l} Mean\_RR, SDNN, RMSSD, SD_1, \\ \cdots, SD_2, T_{Nodding}, T_{Yawning}, T_{Sighing}, T_{Stretching} \end{array} \right]$. L_i represents the fatigue level, that is, the category of sample X_i, which can be evaluated by FAS, as shown in Tab. I, with a total of 5 levels. The SVM classifier uses Gaussian radial basis function to map data to a feature space, which can be expressed as:

$$y(X) = \sum_i c_i L_i \kappa(X_i, S(i)) + b, \tag{10}$$

where c_i is the Lagrange multiplier, $S(i)$ is the SVM classifier, $\kappa(\cdot)$ is the Gaussian kernel function, and b is the bias.

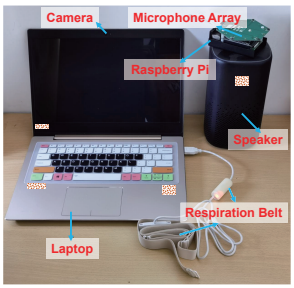


Fig. 9: Experiment setup.

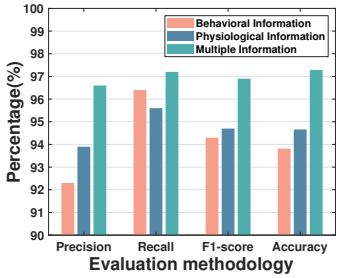


Fig. 10: Overall performance.

IV. IMPLEMENTATION AND EVALUATION

In this section, we introduce the implementation details and provide the evaluation results.

A. Experiment Setup

We implement WakeUp using the experimental devices shown in Fig. 9, including a Raspberry Pi 4B, a regular hexagonal 6-microphone array, an omni-directional speaker, a respiration belt, and a laptop. We use the respiration belt and the laptop camera to record users' real respiration signals and behavioral information, and utilize FAS to assess users' fatigue level. We recruit 20 volunteers (11 males and 9 females) and conduct data collection in four different environments (i.e., dormitory, office, bedroom, and study room) for evaluation. We end up collecting 2000 samples containing fatigue information, of which 60% are used for training and the remaining 40% are used to evaluate WakeUp, and each sample contains a 2-minute signal. On one hand, we segment the 2000 samples to train and evaluate the MobileViT network. On the other hand, we extract the statistical features of the behavioral and physiological information of the 2000 samples to train and evaluate the SVM classifier. All procedures are approved by the Institutional Review Board (IRB) at our institute.

B. Evaluation Methodology

We mainly evaluate WakeUp from the following aspects.

Accuracy. It is the most intuitive performance measure and it is simply a ratio of correctly predicted observations to the total observations.

Precision. It is the ratio of correctly predicted positive observations to the total predicted positive observations.

Recall. It is the ratio of correctly predicted positive observations to the all observations in actual class.

F1-score. It is the harmonic mean of precision and recall.

C. Overall Performance

Overall performance includes three parts: ablation experiment, behavioral information evaluation, and physiological information evaluation.

1) Ablation Experiments: We evaluate the fatigue detection performance of single-information and multi-information through ablation experiments. Fig. 10 shows the fatigue detection performance using only behavioral information, only physiological information, and the fusion of these two kinds of information. It is obvious that the multi-information fusion

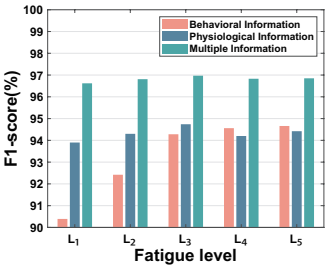


Fig. 11: Detection results of 5 fatigue levels.

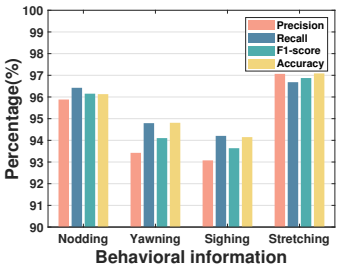
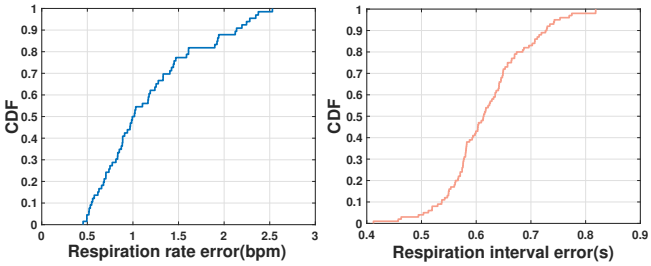


Fig. 12: Classification results of MobileViT network.



(a) Respiration rate (b) Respiration interval

Fig. 13: Error of extracted respiration.

method achieves better detection performance. In particular, the F1-score and accuracy of the multi-information fusion scheme are at least 2.21% and 2.62% higher than those of single-information, respectively, which fully shows the superiority of the multi-information fusion fatigue detection scheme compared with the single information detection scheme.

Fig. 11 presents the detection performance of these three detection schemes for five fatigue levels. When the fatigue degree is low, the F1-score of the detection scheme based on behavioral information is significantly lower, because the occurrence probability of fatigue behavior is relatively low, which can greatly affect the performance of fatigue detection. However, the F1-score of the detection scheme based on physiological information has only a small fluctuation, which also shows the feasibility and reliability of using RRV to detect fatigue. In addition, it is obvious that the multi-information based fatigue detection scheme has excellent detection performance under various fatigue levels, and the F1-score is above 96.62%, which fully reflects the great potential of the multi-information fusion scheme in fatigue detection, and once again proves the effectiveness and robustness of WakeUp.

2) Behavioral Information Evaluation: Fig. 12 shows the classification effect of the MobileViT network on nodding, yawning, sighing, and stretching. It can be seen that the values of these four indicators of all behavioral information are more than 93%, which shows that the MobileViT network has good classification performance and can extract user behavioral information well. It can also reflect the effectiveness of behavioral information feature extraction. In addition, we note that the values of the two indicators of yawning and sighing are lower than those of nodding and stretching. The main reason is that yawning and sighing have a small range of activity, which affects the feature extraction effect of these two behaviors.

3) Physiological Information Evaluation: For respiration signals, the two most important evaluation indicators are respiration rate and respiration interval. We use the error of these two indicators to evaluate the effect of respiration signal extraction. The estimation error of respiration rate is defined as the absolute value of the difference between the estimated respiration rate and the groundtruth. The unit is beats per minute (bpm). The estimation error of respiration interval is defined as the absolute time difference between the estimated respiration interval and the groundtruth, which is an important indicator to measure the accuracy of the boundaries of each respiration. The unit is second. We use the cumulative distribution function (CDF) to quantify these two errors, and the results are shown in Fig. 13. Fig. 13(a) shows the CDF of respiration rate error, with an average error of $1.12bpm$. Fig. 13(b) shows the CDF of respiration interval error, with an average error of $0.61s$. In addition, from the results shown in Fig. 10 and Fig. 11, we can see that the extracted respiration signal and some common features of the calculated RRV are sufficient for the purpose of fatigue detection.

D. Impact of Different Factors

1) Impact of Interference: The user's reflected signal can be affected by the surrounding environment. In III-B, we propose to use microphone array to locate the user and use beamforming to suppress these effects and enhance the user's reflected signal. Now we verify the effectiveness of interference cancellation through experiments. We investigate the impact of people walking around, typing, clicking, talking, and music on WakeUp with and without interference cancellation. The results are shown in Fig. 14. Obviously, in either case, the implementation of interference cancellation is beneficial to improve the performance of WakeUp. In particular, when there are people walking around, no interference cancellation has a huge impact on the performance of WakeUp, the F1-score is only 72.43%, mainly because people have a certain impact on users' behavioral information and physiological information, especially physiological information, WakeUp may not be able to distinguish the respiration signals of users and people around. Typing and clicking have a slight impact on WakeUp. These two small movements may affect the detection performance of yawning and sighing, but these movements are close to the desktop, and we can use interference cancellation technology to remove them. Talking and music have little impact on WakeUp, as WakeUp mainly sends and receives high-frequency signals, and these two sounds mainly contain low-frequency components.

2) Impact of Distance: Next, we evaluate the impact of the distance between the user and the smart speaker on the performance of WakeUp. We increase the distance between the device and the user from $0.3m$ to $2.1m$ every $0.1m$. Fig. 15 shows the F1-score of WakeUp at different distances. When the distance between the user and the device is less than $2.1m$, WakeUp can run well, and its F1-score is more than 95.02%, which can meet the requirements of all environments. In addition, we also find that when the distance is closer

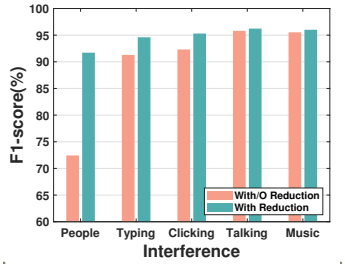


Fig. 14: F1-score under different interferences.

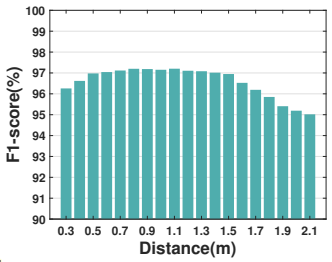


Fig. 15: F1-score under different distances.

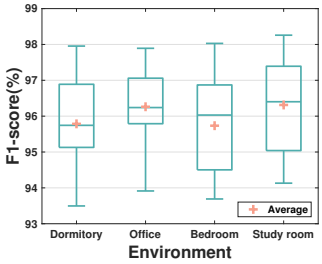


Fig. 16: F1-score under different environments.

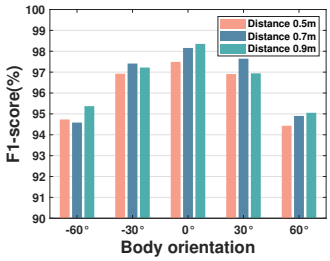


Fig. 17: F1-score under different body orientations.

$(0.3 \sim 0.4m)$, it does not mean that its F1-score can be higher, the main reason is that WakeUp may not be able to capture all the reflected signals of user's behavioral information. When the distance is far $(1.6 \sim 2.1m)$, the F1-score decreases due to weak reflection signal.

3) Impact of Environment: We evaluate the performance of WakeUp under four different environments (i.e., dormitory, office, bedroom, and study room), and the results are shown in Fig. 16. It can be seen that different environments have little impact on WakeUp, with an average F1-score of more than 95.73%. However, it should be noted that since we use line-of-sight signals, there are no obstructions between users and devices when we evaluate the impact of these environments on WakeUp. WakeUp may not work properly when there is an obstruction between the device and the user, which is the limitation of our work.

4) Impact of Body Orientation: When the device is placed in different positions, the user's body orientation relative to the device is different, which may have a certain impact on the performance of WakeUp. Therefore, we study the impact of different body orientations on WakeUp. For convenience, we denote the orientation of the user facing the device as $0°$, the left is $-90° \sim 0°$ area, and the right is $0° \sim 90°$ area. We measure F1-scores of five angles in the $-60° \sim 60°$ area at three distances ($0.5m$, $0.7m$, and $0.9m$), and the results are shown in Fig. 17. WakeUp performs best when the user is at $0°$ relative to the device. When the angle increases from $0°$ to $60°$ (or decreases to $-60°$), the F1-score gradually decreases. But overall, the F1-score is above 94%. In particular, when the angle is greater than $60°$ (or less than $-60°$), the reflected signals from the user's chest and mouth may be difficult to receive, and the performance of WakeUp can be poor. Therefore, we recommend users to place the device at $-60° \sim 60°$ to obtain accurate detection results.

5) Impact of Clothing: In daily life, users may wear various types of clothes, so we evaluate the impact of different clothes on WakeUp. We evaluate the performance of WakeUp under four different dressing scenarios, including T-shirt, sweater, coat, and sweater+coat. The experimental results are shown in Fig. 18. In any case, the average F1-score is greater than 95.16%. However, we also find that when users wear thin clothes (i.e., T-shirt), the F1-score can be higher, while when users wear thick clothes (i.e., sweater+coat), the performance of WakeUp can be slightly affected. We believe that the

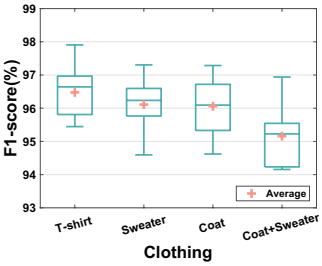


Fig. 18: F1-score under different clothing.

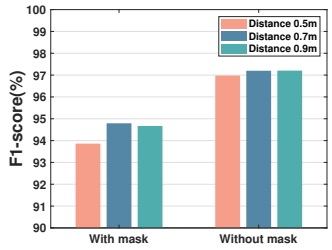


Fig. 19: F1-score with/without mask.

possible reason for the performance degradation is that thicker clothes can attenuate the signal more, so the received chest reflection signal is weak, which can affect the effect of physiological information extraction.

6) Impact of Mask: Many people wear masks during the COVID-19 pandemic. For this reason, we also evaluate the impact on WakeUp with and without masks, and the results are shown in Fig. 19. Wearing a mask has a certain impact on WakeUp, and the F1-score can drop by 2.41% to 3.12%. The main reason is that wearing a mask may make it difficult to receive the reflected signal of the mouth, which can affect the detection accuracy of yawning and sighing, and in turn, affect the performance of WakeUp. But in general, even when wearing masks, the F1-score is still more than 93%, so WakeUp can still work reliably.

V. CONCLUSION

In this paper, we propose WakeUp, a non-contact system that fuses multiple information to detect fatigue in a fine-grained way, to improve people's health levels and work efficiency. WakeUp uses microphone arrays to locate users and adopts beamforming to eliminate interference from the surrounding environment. It can also use the transmitted and reflected FMCW signals to obtain IF signals and phase changes, so as to obtain the user's behavioral information and physiological information at the same time. WakeUp fuses these two kinds of information by extracting the statistical features and combines an SVM classifier to carry out fine-grained fatigue level detection. Extensive experiments show that WakeUp can detect fatigue with an accuracy of 97.28%. We also prove the effectiveness and robustness of WakeUp under different experimental settings.

REFERENCES

[1] M. Tanaka, S. Tajima, K. Mizuno, A. Ishii, Y. Konishi, T. Miike, and Y. Watanabe, "Frontier studies on fatigue, autonomic nerve dysfunction, and sleep-rhythm disorder," *The Journal of Physiological Sciences*, vol. 65, no. 6, pp. 483–498, 2015.

[2] K. Parker, J. M. Horowitz, and R. Minkin, "Covid-19 pandemic continues to reshape work in america," 2022.

[3] Apollotechnical Engineered Talent Solutions, "STARTLING REMOTE WORK BURNOUT STATISTICS," https://www.apollotechnical.com/remote-work-burnout-statistics/, 2022.

[4] T. Zhang, D. Gerlowski, and Z. Acs, "Working from home: small business performance and the covid-19 pandemic," *Small business economics*, vol. 58, no. 2, pp. 611–636, 2022.

[5] J. A. Ricci, E. Chee, A. L. Lorandeau, and J. Berger, "Fatigue in the us workforce: prevalence and implications for lost productive work time," *Journal of occupational and environmental medicine*, pp. 1–10, 2007.

[6] H. J. Michielsen, J. De Vries, G. L. Van Heck, F. J. Van de Vijver, and K. Sijtsma, "Examination of the dimensionality of fatigue: The construction of the fatigue assessment scale (fas)." *European Journal of Psychological Assessment*, vol. 20, no. 1, p. 39, 2004.

[7] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, "Validation of the karolinska sleepiness scale against performance and eeg variables," *Clinical neurophysiology*, vol. 117, no. 7, pp. 1574–1581, 2006.

[8] J. Huang, "Study of driving fatigue based on typical parameters of physiological and operating behavioral characteristics," Master's thesis, South China University of Technology, 2016.

[9] G. Li and W.-Y. Chung, "Estimation of eye closure degree using eeg sensors and its application in driver drowsiness detection," *Sensors*, vol. 14, no. 9, pp. 17 491–17 515, 2014.

[10] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, "Detection of driver's drowsiness by means of hrv analysis," in *2011 Computing in Cardiology*. IEEE, 2011, pp. 89–92.

[11] S. Abtahi, S. Shirmohammadi, B. Hariri, D. Laroche, and L. Martel, "A yawning measurement method using embedded smart cameras," in *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2013, pp. 1605–1608.

[12] S. Hachisuka, "Human and vehicle-driver drowsiness detection by facial expression," in *2013 International Conference on Biometrics and Kansei Engineering*. IEEE, 2013, pp. 320–326.

[13] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 42–55.

[14] Y.-T. Wang, J. Li, R. Zheng, and D. Zhao, "Arabis: An asynchronous acoustic indoor positioning system for mobile devices," in *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2017, pp. 1–8.

[15] K. Sun, T. Zhao, W. Wang, and L. Xie, "Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 591–605.

[16] Y. Xie, F. Li, Y. Wu, and Y. Wang, "Hearfit: Fitness monitoring on smart speakers via active acoustic sensing," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[17] Canalys, "Global smart speaker market 2021 forecast," https://www.canalys.com/newsroom/canalys-global-smart-speaker-market-2021-forecast, 2021.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[19] F. Zhang, Z. Wang, B. Jin, J. Xiong, and D. Zhang, "Your smart speaker can" hear" your heartbeat!" *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–24, 2020.

[20] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 139, no. 5. IET, 1992, pp. 343–350.

[21] G. Yu, T. Lin, Z. Wang, and Y. Li, "Time-reassigned multisynchrosqueezing transform for bearing fault diagnosis of rotating machinery," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1486–1496, 2020.

[22] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[23] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[24] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[25] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2013.

[26] M. V. Yeo, X. Li, K. Shen, and E. P. Wilder-Smith, "Can svm be used for automatic eeg detection of drowsiness during car driving?" *Safety Science*, vol. 47, no. 1, pp. 115–124, 2009.

[27] C. Katsis, N. Ntouvas, C. Bafas, and D. Fotiadis, "Assessment of muscle fatigue during driving using surface emg," in *Proceedings of the IASTED international conference on biomedical engineering*, vol. 262, 2004.

[28] C. Yang, X. Wang, and S. Mao, "Rfid-based driving fatigue detection," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[29] L. Geng, F. Yuan, Z. Xiao *et al.*, "Driver fatigue detection method based on facial behavior analysis," *Comput. Eng*, vol. 44, no. 1, pp. 274–279, 2018.

[30] J. Liu, K. Zhang, W. He, J. Ma, L. Peng, and T. Zheng, "Non-contact human fatigue assessment system based on millimeter wave radar," in *2021 IEEE 4th International Conference on Electronics Technology (ICET)*. IEEE, 2021, pp. 173–177.

[31] W. Jia, H. Peng, N. Ruan, Z. Tang, and W. Zhao, "Wifind: Driver fatigue detection with fine-grained wi-fi signal features," *IEEE Transactions on Big Data*, vol. 6, no. 2, pp. 269–282, 2018.

[32] N. A. B. Amirudin, N. Saad, S. S. A. Ali, and S. H. Adil, "Detection and analysis of driver drowsiness," in *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*. IEEE, 2018, pp. 1–9.

[33] Q. Abbas, "Hybridfatigue: A real-time driver drowsiness detection using hybrid features and transfer learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, 2020.

[34] C. Craye, A. Rashwan, M. S. Kamel, and F. Karray, "A multi-modal driver fatigue and distraction assessment system," *International Journal of Intelligent Transportation Systems Research*, vol. 14, no. 3, pp. 173–194, 2016.

[35] H. Zhang, W. Du, P. Zhou, M. Li, and P. Mohapatra, "Dopenc: Acoustic-based encounter profiling using smartphones," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, 2016, pp. 294–307.

[36] D. He, H. Cao, S. Wang, and X. Chen, "Time-reassigned synchrosqueezing transform: The algorithm and its applications in mechanical signal processing," *Mechanical Systems and Signal Processing*, vol. 117, pp. 255–279, 2019.

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[39] J. Solaz, J. Laparra-Hernández, D. Bande, N. Rodríguez, S. Veleff, J. Gerpe, and E. Medina, "Drowsiness detection based on the analysis of breathing rate obtained from real-time image recognition," *Transportation research procedia*, vol. 14, pp. 3867–3876, 2016.

[40] R. Soni and M. Muniyandi, "Breath rate variability: a novel measure to study the meditation effects," *International journal of yoga*, vol. 12, no. 1, p. 45, 2019.