PDVocal: Towards Privacy-preserving Parkinson's Disease Detection using Non-speech Body Sounds*

Hanbin Zhang
University at Buffalo, SUNY
Buffalo, New York
hanbinzh@buffalo.edu

Chen Song
University at Buffalo, SUNY
Buffalo, New York
csong5@buffalo.edu

Aosen Wang
University at Buffalo, SUNY
Buffalo, New York
aosenwan@buffalo.edu

Chenhan Xu
University at Buffalo, SUNY
Buffalo, New York
chenhanx@buffalo.edu

Dongmei Li
University of Rochester Medical Center
Rochester, New York
dongmei_li@urmc.rochester.edu

Wenyao Xu
University at Buffalo, SUNY
Buffalo, New York
wenyaoxu@buffalo.edu

ABSTRACT

Parkinson's disease (PD) is a chronic neurodegenerative disorder resulting from the progressive loss of dopaminergic nerve cells. People with PD usually demonstrate deficits in performing basic daily activities, and the relevant annual social cost can reach about $25 billion in the United States. Early detection of PD plays an important role in symptom relief and improvement in performance of activities in daily life (ADL), which eventually reduces societal and economic burden. However, conventional PD detection methods are inconvenient in daily life (*e.g.*, requiring users to wear sensors). To overcome this challenge, we propose and identify the non-speech body sounds as the new PD biomarker, and utilize the data in smartphone usage to realize the passive PD detection in daily life without interrupting the user. Specifically, we present *PDVocal*, an end-to-end smartphone-based privacy-preserving system towards early PD detection. *PDVocal* can passively recognize the PD digital biomarkers in the voice data during daily phone conversation. At the user end, *PDVocal* filters the audio stream and only extracts the non-speech body sounds (*e.g.*, breathing, clearing throat and swallowing) which contain no privacy-sensitive content. At the cloud end, *PDVocal* analyzes the body sounds of interest and assesses the health condition using a customized residual network. For the sake of reliability in real-world PD detection, we investigate the method of the performance optimizer including an opportunistic learning knob and a long-term tracking protocol. We evaluate our proposed *PDVocal* on a collected dataset from 890 participants and real-life conversations from publicly available data sources. Results indicate that non-speech body sounds are a promising digital biomarker for privacy-preserving PD detection in daily life.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*.

KEYWORDS

Mobile Health, Parkinson's Disease, Acoustic Sensing.

1 INTRODUCTION

Parkinson's disease (PD), as the second most prevalent neurodegenerative disorder in the world [1], broadly affects 1% of the elderly after 60 and 3% of the elderly after 80 in the U.S. [2]. Early diagnosis and treatment of PD can effectively slow or halt disease progression [3] and extend lifespans (20 or more years) [4]. However, individuals are rarely aware of the early signs of PD in daily life because the initial stage of PD only presents mild or unnoticeable non-motor symptoms (*e.g.,* mood disorders, sleep disorders and voice disorders) [5]. Consequently, they usually miss early treatment and do not seek clinical diagnosis until the mid-stage, by which around 70% [6] of all dopamine neurons may have been permanently impaired. One typical example is the U.S. boxer, Muhammad Ali, who did not receive any treatment until the disease was mid-stage, 4 years after PD onset, which worsened disability complications [7]. This highlights the importance of early detection for which *PDVocal* can improve.

Although the accurate detection of PD has been intensively investigated in clinical medicine and has a rich set of proven approaches (*e.g.*, blood test [8] and neuroimaging abnormalities [9]), the facilities of clinical diagnosis are expensive and not conveniently accessible in either daily life or primary care places. To address this limitation, researchers have explored a set of technologies for PD detection using cost-effective sensors for daily-life applications. For example, Arora *et al.* [10] propose a wearable accelerometer-based

*This work is a pre-print version to appear at MobiCom 2019.

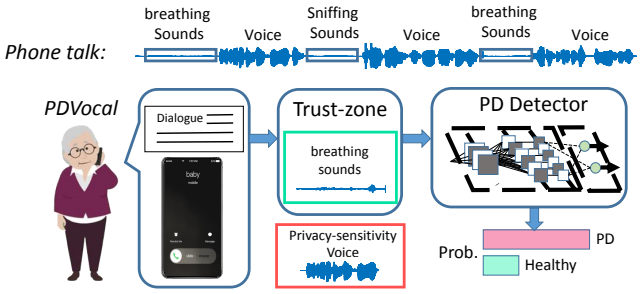


Figure 1: *PDVocal* extracts non-speech body sounds from a user's phone usage (*e.g.,* calling, voice message, voice mailbox and voice chatting) in daily life to perform the pervasive PD risk estimation.

system to extract free balance abnormalities in the time-up-and-go test to identify PD. Several researchers [11, 12] propose using a voice-based system for PD detection. The voice is recorded in a laboratorial environment. To extract the values of fundamental frequency, the subjects are asked to produce vowel [a] at usual intensity for an average period of 10 seconds. Although these techniques are promising to bring an affordable solution to early detection of PD, they require users at risk to be cooperative to perform these tests in their daily routines, and reports from several recent studies [13] show low user adherence in practice. Therefore, the aforementioned methods for PD detection are not practical in daily life. Afterward, considering vocal impairment has been proved as one of the earliest non-motor indicators of PD, a preliminary study, led by Little *et al.* [14], discover that it is possible to extract early signs of PD from voice through phone conversations in daily life. However, their approach can only provide an accuracy of 64%, and such a tel-monitoring approach impairs user privacy. Thereby, positive feedback in user adherence [13] encourages us to explore a vocal-based PD detection system more suited for daily-life usage. Specifically, three key challenges remain as follows: ***(1) Passive Sensing:*** Continuous monitoring using mobile devices removes the daily burden of repeated vocal tasks with current detection systems. ***(2) Privacy-preserving:*** With voice monitoring, we can filter privacy-sensitive content (*e.g.*, personal subjects and information) to enhance privacy and ensure a more trust-worthy method. ***(3) Reliable Detection:*** Considering the daily-life environment, we need a PD detector that is resilient to use conditions and background noise to achieve reliable PD detection.

To satisfy these challenges, we propose *PDVocal* (Fig. 1). *PDVocal* extracts ***passive non-speech body sounds*** from a user's phone usage (*e.g.,* calling, voice message, voice mailbox and voice chatting) in daily life to perform the pervasive PD risk estimation. As described in Section 2, voice has a strong relationship with PD because its generation relies on the cooperation of physical vocal organs. When PD affects these physical vocal organs, the voice alters and we can estimate if a person is PD onset or not by measuring this minor variation. The non-speech body sounds form in the same way. Therefore, non-speech body sounds can also reflect the conditions of these vocal organs. Most of the existing works, however, focus on studying either sustained vowels or free speech for PD detection, and these existing features are not directly applicable to non-speech body sounds, which are usually soft, short and break. To address this problem, we design and implement *ParkinsonNet*, a customized residual Convolutional Neural Network (CNN) to automatically select features and perform a PD risk estimation. We also implement a performance optimizer including an opportunistic learning knob and a long-term tracking protocol to improve the performance of PD detection in daily-life usage.

We evaluate *PDVocal* on a collected smartphone dataset containing 13941 samples (7039 samples from healthy people and 6902 samples from PD patients) contributed by 890 subjects. Our experiments show the performance of *PDVocal* is on par with the state-of-the-art voice-based PD detection approach, which requires subjects to perform tests proactively. Results show that we achieve an average accuracy of 83.3% with the non-speech body sounds which is comparable with the accuracy using state-of-the-art methods (*e.g.,* 85.8% with speech data). We also evaluate the *PDVocal* on publicly available social media data. We extract and label 167 samples of non-speech body sounds from 32-minute-long YouTube videos, and we achieve an average accuracy of 74.7% in PD detection. Our results prove that *PDVocal* is a promising and feasible system to empower the privacy-preserving and high user adherence to PD detection in daily life.

The contribution of our research can be summarized as three-fold:

- To our best knowledge, our work is the first to identify non-speech body sounds as effective indicators of early PD signs. We carry out an in-depth analysis of the interrelation between PD and non-speech body sounds.
- We design a privacy-preserving and passive-sensing system to enable monitoring and estimating the risk of PD in daily life. The foundation of the system rests on the progressive nature of PD and our proposed privacy-isolation technology. Moreover, we implement a dynamic opportunistic learning knob, a PD detector, and a performance optimizer accordingly to enhance the prediction.
- We evaluate our proposed system on the collected dataset and real-life conversations from publicly available

audio sources, such as YouTube. Our results reveal subtle non-speech body sounds collected in an uncontrolled environment can achieve almost the same performance as the professional vocal test in the PD detection. This discovery paves the way for a new approach to PD detection in daily life and other related healthcare areas.

2 BACKGROUND

In this section, we introduce the background on early biomarkers of PD and the impact of PD on vocal organs.

Early biomarkers of PD. Motor symptoms, such as tremor and balance loss, do help detect PD patients from healthy subjects. However, clinical results reveal that those motor symptoms can emerge later at which point about 70% neurons have been permanently impaired [6]. Other researchers observe that some non-motor symptoms can emerge earlier than the motor symptoms. We obtain this conclusion based on the observation of nervous system disorders. Compared with disorders of the central nervous system (CNS) which affect motor symptoms, disorders of the peripheral nervous system (PNS) can emerge earlier causing many of the non-motor symptoms to develop in the early stages [15]. As a representative example, vocal impairment caused by dysfunctions of vocal organs (*e.g.*, lungs and larynx) can reveal early signs of PD better in contrast to other symptoms caused by motor disorders, and therefore, has become a focus area attracting more recent attention [16–19].

Impact of PD on voice. We describe the process of how PD progressively affects the voice in Fig. 2. To begin with, the death of cells happens in the substantia nigra area. It destroys the dopamine pathway and results in insufficient dopamine in these areas. Then, insufficient dopamine induces disorder of the vocal organs (*e.g.*, lung, vocal fold, oral cavity, and nasal cavity). Finally, the disorder of the vocal organs will cause a series of symptoms, such as upper airway obstruction, difficulty in speaking, difficulty in swallowing, excessive salivation, and soft voice.

Conventional tests for PD detection. Since voice alteration is difficult to identify by human experience, researchers have proposed 4 types of tests [20] to understand how voice changes. (1) Sustained vowels: participants are asked to phonate a vowel for several seconds; (2) Diadochokinesia task: participants are required to phonate the occlusive consonants like /pa/-/ta/-/ka/ repeatedly; (3) Reading: participants need to read a specific article; (4) Free speech: participants are asked to talk with another. After collecting data from these tests, researchers then estimate the severity of the PD by analyzing the specific features extracted from the audio data.

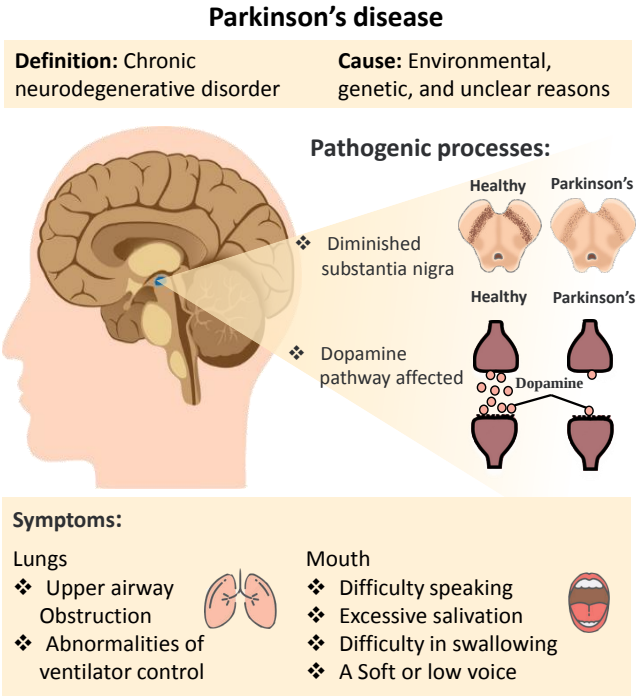


Figure 2: Onset and progressive voice alters due to PD onset and progression. The diminished substantia nigra can progressively influence the vocal organ and finally change the voice.

Although these tests can help PD detection, they show limitations in daily life. First, these existing methods rely on the users to be *cooperative* thus impair user adherence. Second, these approaches are limited by environmental conditions and equipment, which are not accessible in daily life. Third, a method that analyzes free speech can even impair user privacy. These drawbacks encourage us to design a PD detection system that facilitates privacy-persevering and high user adherence.

3 DESIGN CONSIDERATIONS

In this section, we propose our design goals, and we investigate the rationale of employing non-speech body sounds on PD detection.

3.1 Design Goals

We have taken into account the following aspects to facilitate early detection of PD in daily life.

(1) Passive sensing: PD is a progressive disease, and its symptoms grow gradually. To capture the minor variations in the early stages, it is non-trivial to develop a long-term

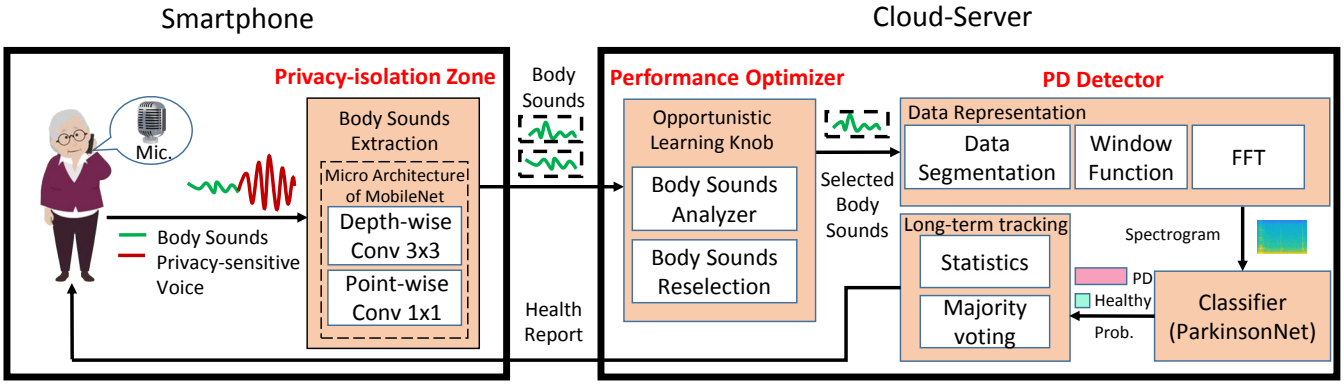


Figure 3: The proposed *PDVocal* framework includes the smartphone end and the server end. The smartphone module consists of the privacy-isolation zone to extract the non-speech body sounds. The server end provides the service of PD detection and user feedback.

monitoring system for PD detection without impairing user adherence.

(2) Privacy-preserving: Privacy-preserving is another essential factor for any long-term monitoring system as no user is willing to expose his or her daily activities. By guaranteeing user privacy, we can achieve user adherence and compliance when performing passive daily monitoring.

(3) Reliable detection: Most existing approaches rely on an experimental environment where the environmental noise is strictly controlled. However, these conditions are no longer consistent at either home or work. In our case, an available approach should be resilient to background noise and user conditions, irrespective of the surrounding environmental conditions.

3.2 A New Digital Biomaker of PD

We explore the solution of PD detection which satisfies the design goals we have mentioned.

Non-speech body sounds. We reasonably infer that non-speech body sounds can reflect PD symptoms. Phonation is a very complicated procedure requiring the coordination of these vocal organs. The lung controls the movement of the air flow, and the friction between vocal organs and the air flow generates the voice. When affected by PD, hypokinesia appearing in the lung, larynx and related vocal organs (*e.g.*, lip and tongue) influences such coordination thus altering the voice. The non-speech body sounds are generated similarly. They are generated by the friction caused by the air flow from the lungs through vocal organs to the mouth and nasal cavity [21]. Therefore, these non-speech body sounds are also highly indicative of the conditions of our lungs [22] and are indicators of PD. This relationship encourages us to start our exploration.

Table 1: A comparison of different approaches for PD detection.

Sensor Type	Sensing Modality	Is Early Biomarker	Is Privacy Preserving	User Adherence
Acce.&Gyro.	Gait	●	○	Low
Webcam	Gait	●	●	High
Mic.	Free Speech	○	●	Low
	Sustain Vowels	○	○	Low
	Body Sounds	○	○	High

○ = Yes ● = No

Advantage and significance. We compare our approach with the existing ones in Table 1. Although accelerometer [23] and webcam [24] can assist gait analysis well, these motor symptoms usually do not appear until disease mid-stage. To detect vocal impairment, free speech and sustained vowels [20] are two current state-of-the-art approaches. However, they either impair user privacy or user adherence, thus showing limitations in daily usage. In contrast to these methods, non-speech body sounds have the properties of being privacy-preserving and can be passively collected in daily life.

4 SYSTEM OVERVIEW

According to the aforementioned design considerations, we design and implement *PDVocal* (Fig. 3), which contains a smartphone end and a server end. The smartphone module consists of a privacy-isolation zone to extract non-speech body sounds. The server end consists of a performance optimizer and a PD detector to provide the service of PD detection and user feedback.

Privacy-isolation zone. The privacy-isolation zone can recognize the composition of an up-coming audio segment. Then, it filters the privacy-sensitive content and only the non-speech body sounds, such as breathing sounds, clearing throat sounds and swallowing sounds, will be further transmitted to the server.

PD detector. We implement a PD detector including the data representation module and the *ParkinsonNet* at the server end to perform the PD risk estimation. To begin with, data representation with spectrogram is adopted for two purposes. (1) It enhances the features in both the time domain and the frequency domain. (2) It transforms the original one-dimension time data to the three-dimension data for the input of the deep neural network. Next, *ParkinsonNet* finishes the automatic features selection and the PD risk estimation.

Performance optimizer. To improve PD detection, we implement a performance optimizer including an opportunistic learning knob and long-term tracking protocol at the server end. The opportunistic learning knob can analyze and remove the outliers. The long-term tracking protocol then associates with the PD detector to provide health report feedback.

As a result, *PDVocal* is working while other smartphone applications, such as calling, voice message and voice chatting, are happening. It does not require the user to be cooperative but can still passively sense the PD-sensitive biomarkers to achieve early detection of PD.

5 PRIVACY-ISOLATION ZONE IN SMARTPHONE

In this section, we introduce the design and implementation of our privacy-isolation zone to extract non-speech body sounds from audio stream at the smartphone end.

Challenges. Given the audio stream, we are motivated to filter out the privacy-sensitive content and only focus on the privacy-irrelevant non-speech body sounds to facilitate a privacy-preserving PD detection. One solution is to train a body-sound recognition model similar to speech recognition [25–27]. Due to the excessive resource cost for the mobile devices, this complicated model is usually deployed at the cloud-server. When the system works, the cloud server receives the data from the mobile end (*e.g.*, smartphone) and feeds back results of the inferences to the users. Therefore, this solution cannot satisfy the privacy-preserving requirement. Another solution is to add extra hardware to support data collection and computing [21], but it can impair user experience.

Solution with MobileNets. To address this problem, we are seeking a solution to embed a small but efficient model for non-speech body sounds extraction at the smartphone end with low system overhead. For this purpose, we adopt MobileNets [28], which are computationally efficient CNN architectures designed specifically for mobile devices. MobileNets adopt the plain architecture which uses depth-wise separable convolutions to build a lightweight deep neural network. In this way, the overhead of MobileNets is low, which can be parameterized to meet the resource constraints of the smartphones in daily-life usage.

Implementation. Specifically, we implement the architecture of *0.25-MobileNet-224* on the smartphone, where the number of 0.25 is the coefficient of the width of the network and the number of 224 indicates the input resolution. The total number of multiply-accumulate operations (MACS) and parameters of *0.25-MobileNet-224* are 41 million and 0.47 million, respectively, which are about 300 times smaller than *VGG-16* [29], a state-of-the-art plain architecture.

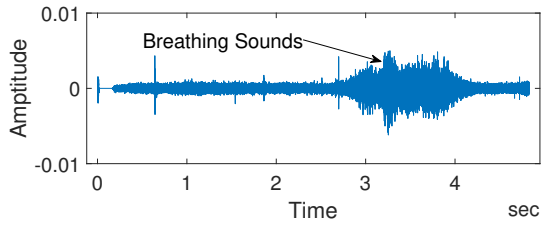
6 PD DETECTOR

In this section, we introduce the core part of our *PDVocal* system, including the data representation module and the *ParkinsonNet*.

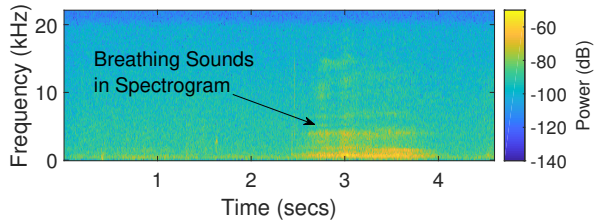
6.1 Data Representation to Augment Features

Although researchers have studied the PD related features of voice for several years, few references are available when referring to non-speech body sounds. Considering that non-speech body sounds are usually low and soft, we are motivated to represent the related features that benefit PD detection. Intuitively, Fast Fourier Transform (FFT) is considered to be one of the state-of-the-art tools to analyze audio signals. However, non-speech body sounds are non-stationary signals thus a simple FFT operation cannot reflect its time-domain features.

Our solution is to divide the non-speech body sounds into several segments by assuming the signal in each segment is stationary, and then we perform the operation of FFT for each segment. In this way, we represent our data in both time domain and frequency domain. To compensate for the truncation effect, the window function is further applied to reduce the spectrum leakage and to improve the spectral resolution. Meanwhile, we adopt the overlapping technique to compensate for the magnitude distortion induced by the window function. In detail, we segment the extracted non-speech body sounds into segments with a length of 200*ms* for each and overlapping with 50%. Considering the prior knowledge that PD influences both loudness and intonation of voice, we select the Kaiser window [30] to maximize the magnitude resolution and the frequency resolution. Then,



(a) Breathing sounds waveform



(b) Spectrogram

Figure 4: An example of time domain wave of the breathing sound and its corresponding spectrogram.

we employ the FFT operation to each segment:

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}. \quad (1)$$

Finally, all segments are formed into an image (Fig. 4):

$$\text{spectrogram}\{x(t)\}(m,\omega) \equiv |X(m,\omega)|^2. \quad (2)$$

The X-axis presents the time dimension, and the y-axis presents the frequency dimension. The third dimension shows the amplitude of a particular frequency at a specific time represented by the color.

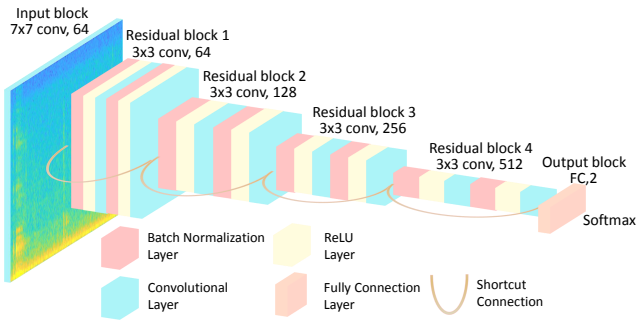


Figure 5: The architecture of the ParkinsonNet. It contains four residual blocks, and each of which contains two convolutional layers.

6.2 ParkinsonNet

Challenges. Although CNN has been viewed as the state-of-the-art tool in a classification problem, it requires a large dataset for training. Otherwise, the network will overfit. The existing state-of-the-art architectures are usually verified in some large datasets (*e.g.*, MNIST [31], CIFAR [32], ImageNet [33], *etc.*) which contain at least 60 thousand samples, and these architectures are unnecessarily large for a small dataset. Therefore, enabling the deep neural network on a small customized dataset is challenging.

Solution with residual architecture. To address this problem, we adopt the residual architecture [34] for the following two aspects. (1) Residual architecture prevents overfitting. The reason is twofold. First, different from other plain networks (*e.g.*,*VGG*), the residual architecture does not contain extra fully connected layers. This property makes it contain few parameters. Second, the residual network contains the batch normalization layers, which help prevent overfitting. This is because the operation of normalization happens on each mini batch, which results in the values of mean and variance being slightly different from one another. This difference can be viewed in that the batch normalization adds some noise to each hidden layer's activations thus generates a slight regularization effect. (2) Residual architecture contains the shortcut connections which help convergence. According to He *et al.* [34], shortcut connections allow the network to learn the identify function better. The identity matrix transmits forward the input data that avoids the data vanishing problem.

Implementation of ParkinsonNet. We design our *ParkinsonNet* by referring to the state-of-the-art residual architecture [34]. To avoid overfitting, we are motivated to remove the neurons to reduce the number of parameters. Notably, we choose to reduce the depth rather than the width to avoid the gradient vanishing. Our final *ParkinsonNet* (Fig. 5) contains 1 input block, 4 residual blocks, and 1 output block. Each residual block contains 2 convolutional layers, and the shortcut connection and pooling are adopted between the residual blocks. The shortcut connection helps connect high-level and low-level features, and pooling helps downsample the feature maps to reduce the spatial size of parameters. The architecture also can be called *ResNet-10* since it contains 10 weighted layers.

7 PERFORMANCE OPTIMIZER

In this section, we develop the opportunistic learning knob and the long-term tracking protocol to optimize PD detection.

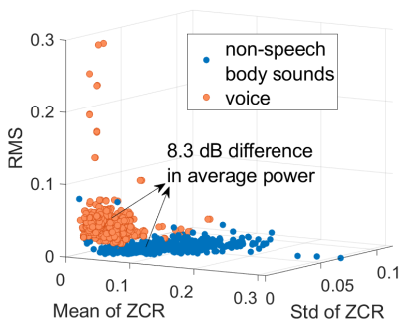


Figure 6: 3D scatter of salient features in comparing non-speech body sounds and voice (plot 500 samples for visualization).

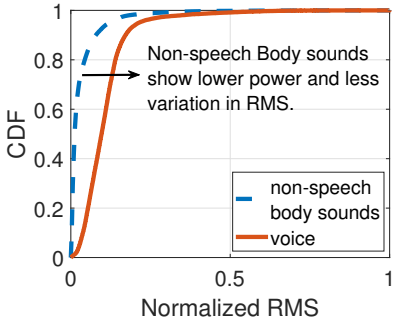


Figure 7: The comparison of RMS value between non-speech body sounds and voice.

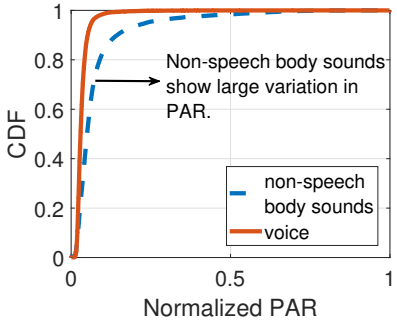


Figure 8: The comparison of PAR value between non-speech body sounds and voice.

7.1 Body Sounds Reselection through the Opportunistic Learning Knob

To improve PD detection, we are motivated to remove the outliers. Our solution is to understand the fundamental properties of non-speech body sounds first. Then, we design an opportunistic learning knob to filter these outliers after the data collection.

Body sounds analysis. We study the power and signal-to-noise ratio (SNR), two fundamental indicators of the audio signal, and we compare these two metrics between the non-speech body sounds and the voice (data description in Section 8.1). Specifically, we adopt the root mean square (RMS) to measure the average power and peak-to-average ratio (PAR) to estimate the SNR. We calculate and present the normalized cumulative distribution function (CDF) graph for each metric in Fig. 7 and Fig. 8, respectively. Our results show the non-speech body sounds present lower power (8.3 dB difference, Fig. 6) and a smaller variation in RMS but present a larger variation in PAR. These results indicate that non-speech body sounds are usually soft and more easily corrupted by environmental factors than the voice.

Opportunistic learning knob. According to the analysis mentioned, we design the opportunistic learning knob to remove the outliers, and we adopt the RMS and PAR as two thresholds for the knob. The workflow can conclude as follows. First of all, we calculate the optimal thresholds of RMS and PAR through brute-force searching in the training. Then, in the inference phase, we filter the collected data with these two thresholds. These two thresholds update when the server progressively collects more data, or the model of *ParkinsonNet* updates.

7.2 Long-term Progression Tracking

Because PD is a progressive neurodegenerative disease where the symptoms are barely noticeable at first but appear gradually over time, proper data accumulation can help better assess the users' condition and generate an accurate prediction. As mentioned, existing proactive methods cannot achieve this goal since users commonly demonstrate low compliance when they are required to perform the repetitive daily test. Our solution can resolve this issue by taking advantage of the passive and unobtrusive sensing in daily smartphone usage. By synthesizing the long-term monitoring results, we can provide users with a more reliable prediction.

In particular, we implement majority voting based on the idea that the long-term assessment can better represent the long-run cumulative effects of PD over the users at risk. We generate one prediction for each extracted body-sound sample. Given a subject n, the prediction for his mth sample can be formulated as:

$$p_{m,n} = \begin{cases} 0, & \text{Health} \\ 1, & \text{PD} \end{cases}. \tag{3}$$

As a result, the majority voting generates the prediction as:

$$\mathcal{P} = \left\{ P_1, P_2, \cdots, P_N | P_n = \frac{1}{M} \sum_{m=1}^{M} p_{m,n} \right\}, \tag{4}$$

where M and N are the number of total body sound samples and the number of subjects, respectively.

8 EVALUATION

In this section, we evaluate *PDVocal* on our collected smartphone dataset.

8.1 Experimental Setup

Participant recruitment. Our study is approved by the Western Institutional Review Board (WIRB). In a 3-month study, we online enroll 890 participants to join our research. We survey the basic demographics for our participants (see Table 2). Among all the people, 567 are male (64%) and 323 are female (36%). All the participants come from the U.S., and they are from 60 years old to 85 years old. To obtain the ground truth, we cooperate with the professional medical institute to evaluate every subject with the Unified Parkinson's Disease Rating Scale (UPDRS), a standard clinical diagnosis for Parkinson's disease. Each subject requires taking a series of tests, such as speech test, facial expression test, hand movement test and gait analysis. Each subject will receive a score when finishing each test, and the physician diagnoses the subject as PD or non-PD according to their performance. As a result, 321 subjects are professionally diagnosed as PD patients, and the remaining 569 subjects are diagnosed as healthy people.

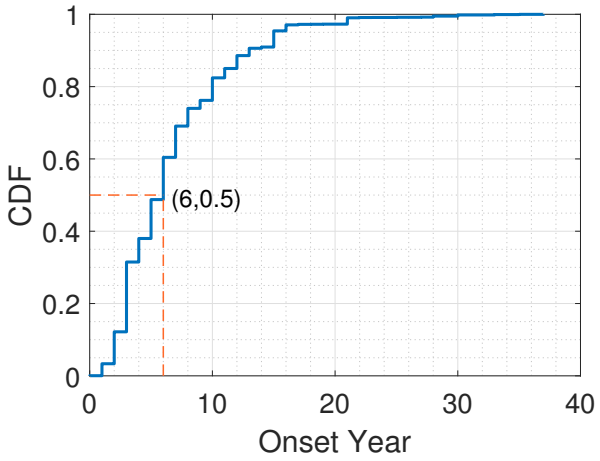


Figure 9: The CDF graph shows the onset information for the PD subjects, 50% of whom are onset longer than 6 years.

We also record the concrete onset dates (Fig. 9) for each PD patients. The age at PD onset is from 1 to 37 years among our participants. The median number is 6, showing there are more than 50% people are PD onset longer than 6 years. Further, there are 31.5% subjects are PD onset less than 3 years, which we then consider them as early-onset patients [35].

Data collection. We extract non-speech body sounds in an uncontrolled daily-life environment. Participants are asked to install our smartphone APP containing a sustained vowel test. It includes a pre-startup phase and a testing phase (Fig. 10). Specifically, the pre-startup phase is a 5-second timeout to remind the start time of the testing phase. And the testing phase is to ask the subjects to perform the sustained vowel [a] for 10 seconds. We turn on the built-in microphone at both the pre-startup phase and the testing phase. Through this method, we can successfully collect the non-speech samples of body sounds in the pre-startup phase. In particular, each onset PD subject is asked to perform the test before any medication. This is to reduce the interference caused by the treatment. Through a 3-month-long experimental period, we collect 13941 samples of body sounds, such as breathing sounds, clearing throat sounds and swallowing sounds, of which 7039 samples are from healthy people and 6902 from PD patients.

Table 2: Demographic survey.

	No. (%)	
Characteristic	**PD**	**Health**
Demographic		
White race	250 (28.0)	552 (62.0)
Higher education	260 (29.2)	575 (64.6)
Male	214 (24.0)	353 (39.7)
Female	107 (12.0)	216 (24.2)
User Condition		
Smoker	119 (13.4)	245 (27.5)
Non-smoker	143 (16.1)	370 (41.6)
iPhone 5/5s/5c	74 (8.3)	223 (25.1)
iPhone 6/6p	181 (20.3)	285 (32.0)

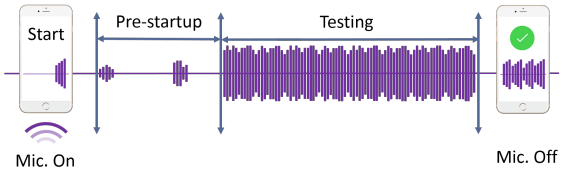


Figure 10: Our data collection protocol includes a pre-startup phase and a testing phase. Pre-startup phase is a timeout to remind the start time of the testing, and we extract the non-speech body sounds from the pre-startup phase.

Control group. Meanwhile, we obtain 13941 samples of voice collected from the testing phase. These sustained vowel samples are the current state-of-the-art materials adopted for PD detection [20]. We set these samples as the control group for comparison in the following evaluations.

8.2 PD Prediction Performance

Evaluation metrics. We use the following metrics that are widely used in mobile health.

- Accuracy: Accuracy describes the fraction of samples that are correctly predicted. It is formulated as $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.
- Precision: Precision is defined as the fraction of predicted PD samples that truly contain PD biomarkers, *i.e.,* $precision = \frac{TP}{TP+FP}$. It measures the robustness of our system against false positives.
- Recall: Recall is defined as the fraction of PD samples that are detected over the total amount of PD samples, *i.e.,* $recall = \frac{TP}{TP+FN}$. It measures our system's ability in detecting all the PD onset people without misses.
- F1-measure: F1 considers both precision and recall, and is computed as the harmonic average of the two, *i.e.,* $f1 = \frac{2 \cdot p \cdot r}{p+r}$.

Dataset split. We adopt the hold-out validation. In each validation, we randomly split our dataset into five parts. We select four of them as training and adopt the remaining one as testing. We ensure the positive and negative samples are balanced in each part.

Implementation. We implement our neural network in PyTorch. We adopt Adam optimizer with an initial learning rate of 0.001. Data augmentation methods including random resize and crop, random horizontal flip, and color jitter are adopted.

Overall performance. We first evaluate our collected dataset on four different models, including the deep neural network and the traditional classifiers. To evaluate our dataset on the traditional machine learning classifiers, we adopt the Mel-frequency cepstral coefficients (MFCCs), a set of widely used features for voice analysis, as the features. To extract MFCCs, we adopt a window size of 2048 with an overlap length of 1024, and we choose 13 coefficients for each window. The data preparation for deep neural network is described in Section 6.

Fig. 11 and Fig. 12 present the accuracy and the f1-measure, respectively. Non-speech body sounds can achieve an accuracy of 83.3% ± 1.6% and a f1-measure score of 82.6% ± 1.1%. These two numbers are 85.8% ± 1.4% and 83.6% ± 1.9% for voice. The difference in performance between non-speech body sounds and voice is close, implying the non-speech body sounds can be viewed as biomarkers for PD detection.

The traditional machine learning model cannot achieve a good performance. The reason is that most of the applied features in traditional machine learning need to be identified by a domain expert to reduce the complexity of the data

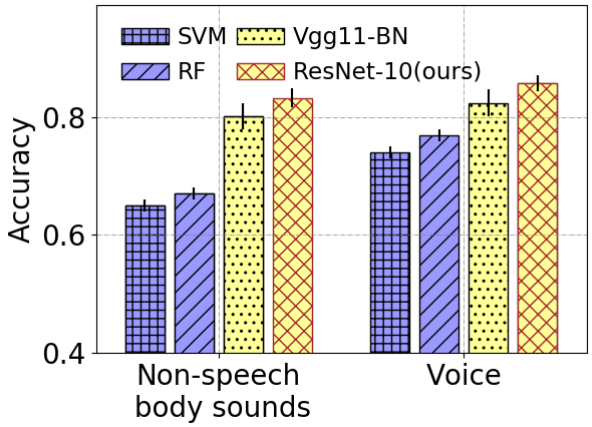


Figure 11: The comparison of the average accuracy between non-speech body sounds and voice.

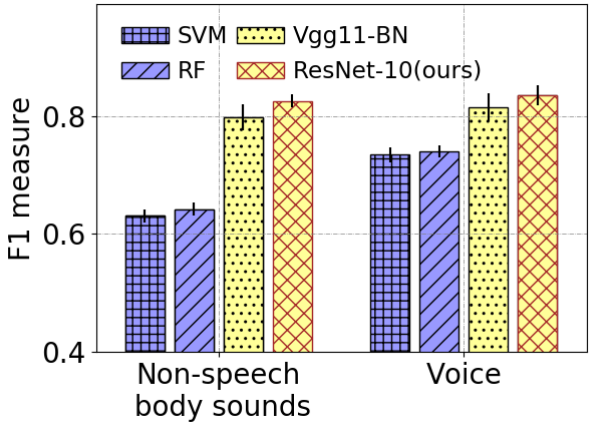


Figure 12: The comparison of the average f1-measure between non-speech body sounds and voice.

and make patterns more visible to learning algorithms to work. In contrast, deep neural network as discussed before is that they try to learn high-level features from data in an incremental manner. This eliminates the need for domain expertise and feature extraction. In our case, MFCCs represent the compressed vocal features. Thereby, they can lose some essential information related to PD.

Fig. 13 plots the normalized confusion matrix. The average recall is 80.2%, and the average precision is 85.2%, respectively. The precision is higher than recall, showing that our system has a lower probability to predict a healthy subject as a PD subject, but correspondingly, a higher probability to predict a PD subject as a healthy subject. The reason is

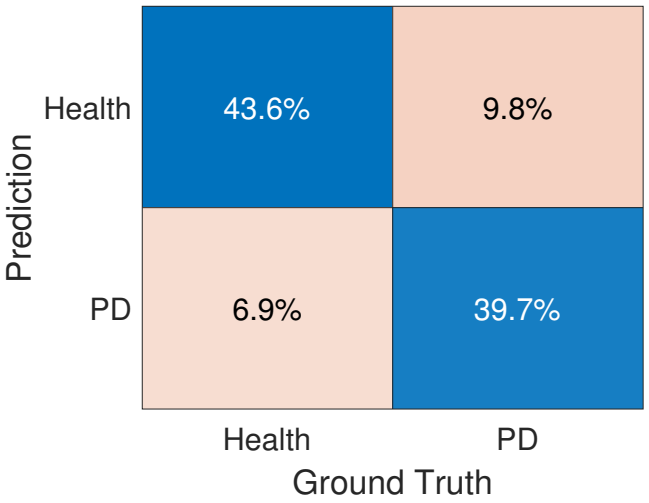


Figure 13: The normalized confusion matrix of PD detection on non-speech body sounds. The precision is higher than recall, showing that *PDVocal* has a lower probability to misclassify a healthy subject as a PD subject, but a higher probability to predict a PD subject as a healthy subject.

twofold. First, symptoms can be different from person to person. We cannot detect PD biomarkers from non-speech body sounds in these subjects whose vocal organs are not impaired by PD. Second, long-term medication can interfere with the PD prediction. In this case, PD biomarkers are not very obvious.

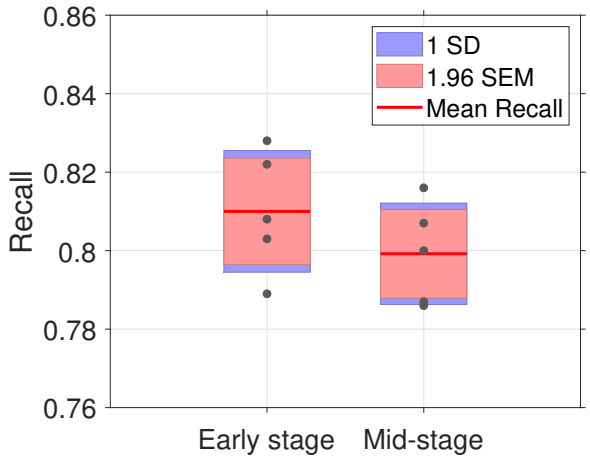


Figure 14: The comparison of recall between the early stage and mid-stage. The results are represented as points, which are layered over a 1.96 SEM in the pink patch and a 1 SD in the blue patch.

Early stage v.s. mid-stage. We further analyze the system performance concerning different PD stages. Generally, PD includes the early stage and the mid-stage [35]. The former one refers to the PD onset less than or equal to 3 years, while the latter one represents the PD onset longer than 3 years. In our testing, there are 26.1% samples that belong to early stage, and 73.9% samples that belong to mid-stage on average.

Fig. 14 shows the comparison of recall between the two stages. The results are represented as points, which are layered over a 1.96 Standard Error of Mean (SEM) in the pink patch and a 1 Standard Deviation (SD) in the blue patch. The average recall is 81.0% for the early stage group and 79.9% for the mid-stage group. The little performance variation between two stages strongly indicates the feasibility of transforming the onset PD detection solution into early PD detection. Moreover, it proves our hypothesis that the non-speech body sounds exist in all stages of PD and can be utilized as the new PD biomarker for unobtrusive and passive daily monitoring. In such a way, we facilitate preventive PD healthcare in daily life.

8.3 The Performance of Privacy Isolation

We evaluate the performance of our privacy-isolation zone at the smartphone end, containing the MobileNet to screen the non-speech body sounds and the voice.

Evaluation metrics. We employ the accuracy and False Positive Rate (FPR) as two metrics to measure performance. Accuracy measures the fraction of samples that are correctly predicted. The FPR describes the fraction of voice samples that are mispredicted as the body sounds samples, *i.e.*, $FPR = \frac{FP}{FP+TN}$. It measures the ability to protect privacy, and as this number should be lower, the better the performance of privacy-isolation.

Dataset and set up. We evaluate our model on our smartphone dataset. We segment each audio sample into segments, and we drop a segment if it only contains background noise. Afterwards, we label the segments from the pre-startup phase as body sounds and the segments from the testing phase as voice, respectively. For evaluation, we adopt the hold-out method which conducts each experiment by 10 runnings with random initialization. The ratio of the training set and testing set is 4:1.

Results. We evaluate the segmentation size of 0.5 second and 1 second, respectively. Table 3 shows that both the 0.5 second and 1 second long segment can achieve a classification accuracy of 99%. Further, the probability to falsely predict a voice sample as body sound is 1.1±0.5% when the segment size is 1 second long, and this probability is 1.0±0.7% when the segment size is 0.5 second long. Results show that

our privacy-isolation zone can well differentiate the non-speech body sounds and voice, and show the potential to protect the privacy-sensitive content in daily life.

Table 3: Accuracy and FPR when the segment size is different.

Segment Size	Accuracy	FPR
0.5 *sec*	98.8±0.4%	1.0±0.7%
1 *sec*	99.0±0.3%	1.1±0.5%

8.4 The Performance of Opportunistic Learning Knob

We evaluate the performance of the opportunistic learning knob. We employ the same evaluation metrics in Section 8.2.

Set up. We set up the following three groups of experiments to evaluate the performance of our opportunistic knob. (1) Group 1: We remove 10% samples with the lowest PAR value in both training and testing; (2) Group 2: We remove the bottom 10% samples with the lowest RMS in both training and testing; (3) Group 3: We remove the bottom 20% samples with the lowest RMS and PAR in both training and testing. We also set up a control group without removing any data.

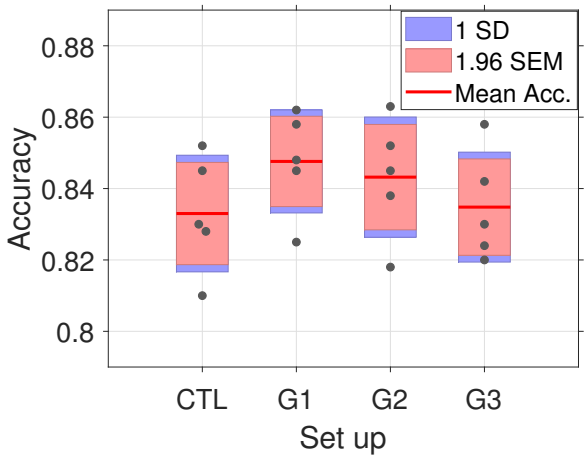


Figure 15: The evaluation for the opportunistic knob. The results are represented as points, which are layered over a 1.96 SEM in the pink patch and a 1 SD in the blue patch.

Results. Fig. 15 shows that removing the outliers can help achieve better performance. The results are represented as points, which are layered over a 1.96 SEM in the pink patch and a 1 SD in the blue patch. Compared to the control group, the average accuracy of Group 1 increases to 84.8%, indicating that daily-life body-sound samples contain environmental noise, and filtering these noised samples can have a positive result. The average accuracy of the Group 2 increases to 84.3%, indicating our knob selects the samples containing prominent biomarkers. Our results indicate that bigger data are not always better data and removing the outliers in some cases improve the performance of PD detection.

8.5 Micro Benchmarks on Diverse Background

We evaluate the influence of diversity on PD detection. We employ the accuracy, precision-recall (PR) curve, and area under the PR curve (AUC-pr) as the metrics.

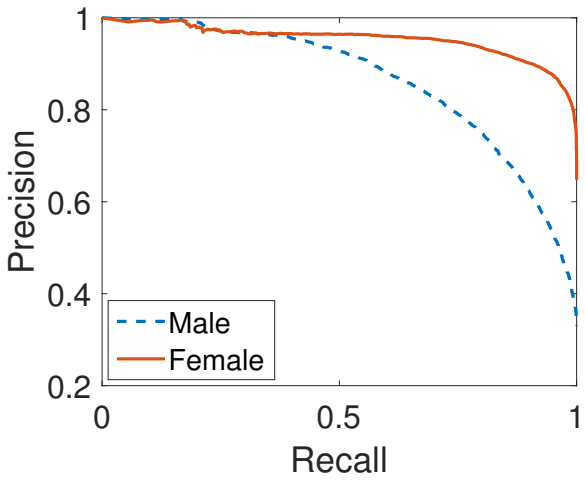


Figure 16: The comparison of PR curves between male and female subjects. The female represents both higher recall and higher precision than the male.

Impact of gender. As males and females have a different construction in vocal organs, we show interests in whether this difference can influence PD detection. We observe that there is a clear advantage for the female (Table 4). First, the average accuracy of the female is 84.2%, which is 1.4% higher than the one of the male. It is also apparent in the PR curve (Fig. 16) where the PR curve of the female can entirely encase the one of the male. Our results show that, in some cases, the female can benefit from these differences and have a more reliable PD detection.

Impact of smoking. Long-term smoking can impair the vocal organs, thus further influence breathing sounds. Thereby, we wish to understand if this impairment can influence the PD detection performance. Our results show that smokers

Table 4: Performance benchmarking.

	Gender		Smoking Survey		Phone Type	
	Male	Female	Smokers	Non-smokers	iPhone 5	iPhone 6
Acc.(%)	82.8	84.2	84.7	82.0	83.7	79.7
AUCpr	0.754	0.827	0.793	0.762	0.889	0.826

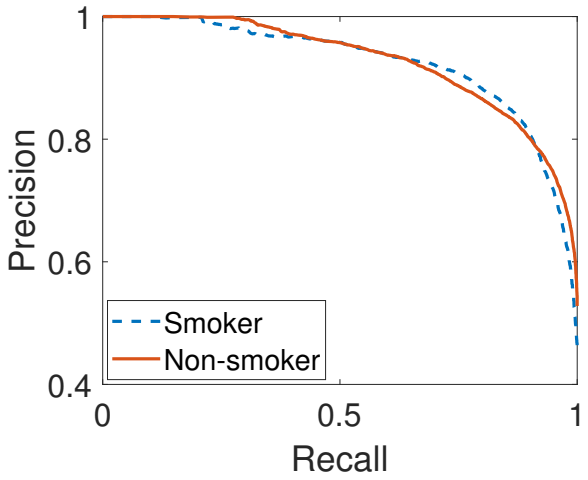


Figure 17: The comparison of PR curves between smokers and non-smokers.

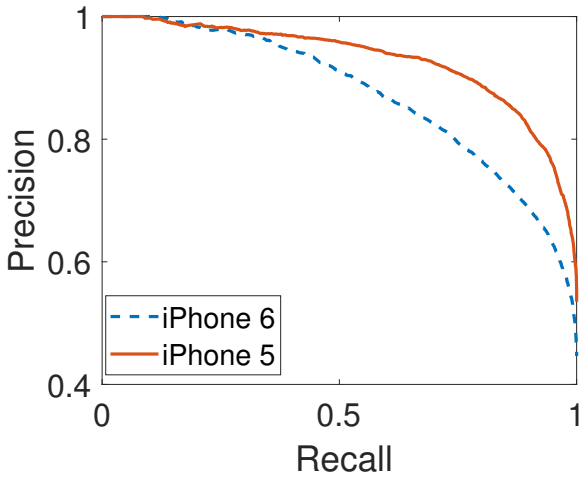


Figure 18: The comparison of PR curves between iPhone 5 and iPhone 6.

can have a more reliable prediction. First, the average accuracy of smokers is 2.7% higher than the non-smokers (Table 4). Second, the AUC-pr is 0.793 for smokers, and this number is 0.762 for non-smokers. Our results indicate that smoking in some cases influence the PD detection.

Impact of smartphone. The position and design of the microphone array and subsequent hardware for different phones are different, and we want to understand if this difference influences PD detection. We observe a huge lead in the performance of the iPhone 5 over the iPhone 6. First, there is a gap of 4.0% in average accuracy (Table 4). Also, the PR curve (Fig. 18) of the iPhone 5 can entirely encase the one of the iPhone 6. This is because the built-in microphone (typically a condenser) is progressively optimized for capturing speech [21]. This change makes the new phones a little bit difficult to capture soft non-speech body sounds.

8.6 System Overhead on Mobile Devices

We evaluate the system overhead at the smartphone end, which contains the privacy-isolation zone to extract the non-speech body sounds from the voice. We implement our model on four different types of smartphones employing the Tensorflow Mobile.

We evaluate latency and runtime, which are the two most significant factors that influence real-time usage. We also evaluate memory and CPU utility to understand the recourse cost of our model. For our objectives, we set up our experiments by continuously running our module for 30 seconds on the smartphones. Meanwhile, we measure the runtime, the memory cost, and the average CPU load.

Table 5: System overhead on different types of smartphones.

Phone Brand	Latency (*ms*)	Throughput	Runtime (*ms*)	Mem. (*MB*)	Aver. CPU(%)
Nexus 5	224.35	338	88.92±13.03	5.14	17.4
Galaxy S7	216.53	508	59.02±7.48	5.14	25.8
Pixel	188.63	519	58.05±7.87	5.14	17.6
Pixel 2	116.89	550	54.49±6.82	5.14	16.1

Table 5 summarizes the system overhead on four different smartphones. Overall, the latency (warmup runtime) can be reduced by improving the capabilities of the smartphone. The latency is 224.35 milliseconds for Nexus 5 (released in 2013),

and this number is 116.89 milliseconds for Pixel 2 (released in 2017). As well as the latency, the runtime is related to the capabilities of the smartphone. In the 30-second-long benchmark, the average runtime for Nexus 5 is 88.92±13.03 milliseconds, and this number is 54.49±6.82 milliseconds for Pixel 2. Our results also show the potential of the privacy-isolation zone in real-life applications. The Nexus 5 can achieve a frame per second (FPS) of 338/30 = 11.3, which can well satisfy the real-time application whose bottleneck is 2 fps for a 0.5-second-long audio segment. Further, the extra memory caused by our module is 5.14 MB. Since the architecture of MobileNet is invariant, this number is consistent for all brands of phones.

9 EVALUATION WITH THE SOCIAL MEDIA DATA

To in-depth explore the availability of PD detection based on the non-speech body sounds, we evaluate *PDVocal* with social media data.

Benchmark preparation. We collect the material from YouTube. Table 6 presents the demographic survey of our dataset. In total, it contains 10 people, of which 5 subjects are clinically diagnosed as PD and the other 5 subjects are healthy people. In particular, the 1st and 2nd subjects are sharing how PD will alter their voice. The 3rd subject is attending a TV program to discuss the experience of PD treatment and recovery. The 4th subject is attending a TV show. The 5th subject is doing a presentation. Meanwhile, all 5 healthy people are doing presentations. Further, we fill the number of age and onset year which are calculated according to the released date of the videos, if users provide them. We convert these videos to audios, and we employ the spectrogram to represent audio data. We directly adopt the pre-trained model in Section 8.2. For each subject, we calculate the number of correctly predicted samples and then calculate the accuracy.

Table 6: Demographic survey.

Video Index	Age	Gender	Onset Year	Video Length	Extracted Samples
1	60	F	23	4:44	32
2	61	F	15	2:43	23
3	53	M	13	4:10	8
4	60	F	23	2:27	5
5	60	F	15	5:16	21
6	28	M	NA	2:55	15
7	24	M	NA	3:38	28
8	25	F	NA	2:53	15
9	20	F	NA	1:07	5
10	38	M	NA	2:39	15

Results with unattended social media data. Our privacy-isolation zone helps extract 167 samples of body sounds from 32-minute-long YouTube videos. It indicates that *PDVocal* can provide PD detection about 5 times every minute. Compared with other test-based approaches, our passive sensing protocol can frequently assess the health conditions and provide more opportunities for early detection of PD in daily life. Fig. 19 shows the accuracy of PD detection for each subject. *PDVocal* achieves an average accuracy of 74.7%. In particular, *PDVocal* can achieve an average accuracy of 91.3% for healthy people, and this accuracy is 58.2% for people with PD. The reason is, however, three-fold. First, our model is trained from a smartphone dataset. Therefore, it can lose generalization ability on the other domain. Second, the recording device can be far away from a subject during the video recording. Therefore, the arriving non-speech sounds can be very weak. In this case, the PD biomarker is not obvious. Third, PD symptoms can be different from person to person. It is difficult to detect PD biomarkers from subjects whose vocal organs are not affected by PD.

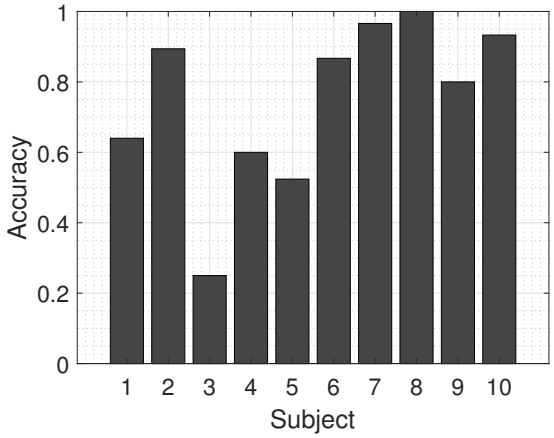


Figure 19: The accuracy of PD detection on 10 YouTube videos.

10 RELATED WORK

Related work falls within three areas, including PD detection methods, mobile health systems, and body sounds and applications.

10.1 PD Detection Methods

We have witnessed great advancement in exploring a variety of technologies for PD detection. Tsanas *et al.* [36] applied voice signal processing algorithm into extracting features from a piece of sustained vowel phonations and mapped the relationship between those features and UPDRS. In 2012,

the same group applied a state-of-the-art machine learning algorithm, SVM and random forest, to detect PD from healthy people in a generated dataset [37]. Shahbakhi *et al.* [38] applied a genetic algorithm (GA) for classification between healthy subjects and PD-afflicted people and Indira *et al.* [39] applied pattern recognition and C-means clustering-based approaches for the screening between healthy and PD-afflicted people in 2013 and 2014, respectively. There is also much work adopting non-voice material into PD detection. Giancardo *et al.* [40] revealed the relationship between key hold time (the time required to press and release a key on a computer keyboard) and PD. Pereira *et al.* [41] studied the relationship between handwritten dynamics and PD with a smartpen. Eskofier *et al.* [42] applied wearable sensors to record the participants' mobility pattern during a specific mobility task.

Our work, however, is different from all the existing work. We observe that low user adherence becomes the first factor in impairing the daily-life PD detection. A passive-sensing system can help, but no existing work has discussed the privacy issue. To address this problem, we propose to extract passive non-speech body sounds in daily life. Compared with other methods, body sounds contain low privacy content (*e.g.*, conversation content or location information), thereby providing a privacy-persevering and passive-sensing protocol.

10.2 Mobile Health Systems

Mobile health (mHealth) is an emerging area of interest for researchers in recent years [43–47]. Lu *et al.* developed Stresssense [48] to recognize stress from the human voice using smartphones. Bui *et al.* developed Pho_2 [49] to measure the blood oxygen level of a person adopting the camera and flashlight on a smartphone. Farhan *et al.* applied data from GPS and accelerometer of the smartphone to perform depression screening [50]. Gao *et al.* developed Healthaware [51], a smartphone-based system utilizes the embedded accelerometer to monitor daily physical activities and the built-in camera to analyze food items to control obesity.

10.3 Body Sounds and Applications

Body sounds related research has attracted much attention in recent years. Yatani *et al.* [52] proposed BodyScope, a wearable neckpiece to capture several kinds of body sounds to predict activity. Hao *et al.* developed iSleep [53] to monitor sleep quality using a smartphone. Larson *et al.* [54] proposed a cough sensing system with a low-cost microphone. Amft and Troster applied a combination of sensors to model eating behavior [55]. Nirjon *et al.* developed Musicalheart [56], a music recommendation system through sensing body vibrations. Chauhan *et al.* developed BreathPrint [57], a new behavioral biometric signature based on audio features derived from an individual's commonplace breathing gestures. Some work also focuses on how to extract body sounds. In BodyBeat [21], the authors designed and implemented a customized sensor system to sense and classify an array of body sounds including eating, drinking, deep breathing, clearing throat, coughing, sniffling and laugh.

Our work observes that the non-speech body sounds contain PD biomarkers. Our system *PDVocal* extracts non-speech body sounds from a user's phone usage in daily life to perform the pervasive PD risk estimation.

11 LIMITATION

We present *PDVocal*, a privacy-preserving and passive sensing mobile system for daily life PD detection. By designing a privacy-isolation zone at the smartphone end and a PD detector in the server end, *PDVocal* can extract the passive non-speech body sounds from daily-life phone usage and perform continuous PD detection.

PDVocal is a closer step towards PD detection in daily life. However, it exhibits some limitations. First, the trained model for body sounds extraction and PD detection is based on our collected smartphone dataset. Performance may degenerate in a new environment on a new subject. This problem can be addressed by collecting more data from more subjects. Second, our work focuses on non-speech body sounds but which types of sound (*e.g.*, breathing sound or swallowing sound) better contribute to the trained model is not clear. In the future work, we will not only improve PD detection performance, but explore a deeper understanding of PD biomarkers in vocal body sounds. Third, our collected dataset does not include young onset-onset patients. We plan to collect more data from a broader sample of subjects in the future. Despite these limitations, we believe our work is an essential step towards PD detection in daily life.

12 CONCLUSIONS

In this paper, we presented, *PDVocal*, the first mobile system that utilizes non-speech body sounds to facilitate the continuous monitoring and estimation of PD risk in daily life. It works by sensing vocal sounds through the built-in microphone, isolation of privacy-sensitive content and detection with a customized convolutional neural network. *PDVocal* demonstrates the significant advantages and is a promising step in the real-world deployment of a privacy-preserving passive-sensing mobile health system towards a large population in the future.

ACKNOWLEDGMENTS

We thank anonymous reviewers and shepherd for their insightful comments on this paper.

REFERENCES

[1] D. Hirtz, D. Thurman, K. Gwinn-Hardy, M. Mohamed, A. Chaudhuri, and R. Zalutsky, "How common are the "common" neurologic disorders?" *Neurology*, vol. 68, no. 5, pp. 326–337, 2007.

[2] E. R. Dorsey and B. R. Bloem, "The parkinson pandemic-a call to action," *JAMA neurology*, vol. 75, no. 1, pp. 9–10, 2018.

[3] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[4] W. Poewe, "The natural history of parkinson's disease," *Journal of neurology*, vol. 253, no. 7, pp. vii2–vii6, 2006.

[5] K. R. Chaudhuri, D. G. Healy, and A. H. Schapira, "Non-motor symptoms of parkinson's disease: diagnosis and management," *The Lancet Neurology*, vol. 5, no. 3, pp. 235–245, 2006.

[6] M. M. Goldenberg, "Medical management of parkinson's disease," *Pharmacy and Therapeutics*, vol. 33, no. 10, p. 590, 2008.

[7] R. L. Brey, "Muhammad ali's message: Keep moving forward," *Neurology Now*, vol. 2, no. 2, p. 8, 2006.

[8] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, *et al.*, "Mutation in the α-synuclein gene identified in families with parkinson's disease," *science*, vol. 276, no. 5321, pp. 2045–2047, 1997.

[9] S. Fahn, "Description of parkinson's disease as a clinical syndrome," *Annals of the New York Academy of Sciences*, vol. 991, no. 1, pp. 1–14, 2003.

[10] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. Biglan, E. Dorsey, and M. Little, "Detecting and monitoring the symptoms of parkinson's disease using smartphones: A pilot study," *Parkinsonism & related disorders*, vol. 21, no. 6, pp. 650–653, 2015.

[11] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, *et al.*, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[12] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[13] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, *et al.*, "The mpower study, parkinson disease mobile data collected using researchkit," *Scientific data*, vol. 3, p. 160011, 2016.

[14] M. Little, "Ubiquitous, inexpensive non-invasive technologies for objective detection and monitoring of parkinson's symptoms," 2013.

[15] F. L. Pagan, "Improving outcomes through early diagnosis of parkinson's disease." *The American journal of managed care*, vol. 18, no. 7 Suppl, pp. S176–82, 2012.

[16] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[17] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[18] J. R. Duffy, *Motor Speech Disorders-E-Book: Substrates, Differential Diagnosis, and Management.* Elsevier Health Sciences, 2013.

[19] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: A longitudinal case study," *Brain and cognition*, vol. 56, no. 1, pp. 24–29, 2004.

[20] L. Jeancolas, H. Benali, B.-E. Benkelfat, G. Mangone, J.-C. Corvol, M. Vidailhet, S. Lehericy, and D. Petrovska-Delacrétaz, "Automatic detection of early stages of parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients," in *Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on.* IEEE, 2017, pp. 1–6.

[21] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "Bodybeat: a mobile system for sensing non-speech body sounds." in *MobiSys*, vol. 14, 2014, pp. 2–13.

[22] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, vol. 2, pp. CCRPM–S530, 2008.

[23] L. Pepa, F. Verdini, M. Capecci, and M. Ceravolo, "Smartphone based freezing of gait detection for parkinsonian patients," in *Consumer Electronics (ICCE), 2015 IEEE International Conference on.* IEEE, 2015, pp. 212–215.

[24] B. Takač, A. Català, D. R. Martín, N. Van Der Aa, W. Chen, and M. Rauterberg, "Position and orientation tracking in a ubiquitous monitoring system for parkinson disease patients with freezing of gait symptom," *JMIR mHealth and uHealth*, vol. 1, no. 2, 2013.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[26] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 8599–8603.

[27] L. Deng, "Achievements and challenges of deep learning-from speech analysis and recognition to language and multimodal processing," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[28] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] J. Kaiser and R. Schafer, "On the use of the i 0-sinh window for spectrum analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 105–107, 1980.

[31] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, 2010.

[32] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* Ieee, 2009, pp. 248–255.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] G. A. Kang, J. M. Bronstein, D. L. Masterman, M. Redelings, J. A. Crum, and B. Ritz, "Clinical characteristics in early parkinson's disease in a central california population-based study," *Movement disorders: official journal of the Movement Disorder Society*, vol. 20, no. 9, pp. 1133–1142, 2005.

[36] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[37] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[38] M. Shahbakhi, D. T. Far, E. Tahami, *et al.*, "Speech analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine," *Journal of Biomedical Science and Engineering*, vol. 7, no. 4, pp. 147–156, 2014.

[39] I. Rustempasic and M. Can, "Diagnosis of parkinson's disease using fuzzy c-means clustering and pattern recognition," *Southeast Europe Journal of Soft Computing*, vol. 2, no. 1, 2013.

[40] L. Giancardo, A. Sanchez-Ferro, T. Arroyo-Gallego, I. Butterworth, C. S. Mendoza, P. Montero, M. Matarazzo, J. A. Obeso, M. L. Gray, and R. S. J. Estépar, "Computer keyboard interaction as an indicator of early parkinson's disease," *Scientific reports*, vol. 6, p. 34468, 2016.

[41] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided parkinson's disease diagnosis from handwritten dynamics," in *Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on*. IEEE, 2016, pp. 340–346.

[42] B. M. Eskofier, S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz, *et al.*, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for parkinson's disease assessment," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 655–658.

[43] K. Sha, G. Zhan, W. Shi, M. Lumley, C. Wiholm, and B. Arnetz, "Spa: a smart phone assisted chronic illness self-management system with participatory sensing," in *Proceedings of the 2nd International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments*. ACM, 2008, p. 5.

[44] T. Denning, A. Andrew, R. Chaudhri, C. Hartung, J. Lester, G. Borriello, and G. Duncan, "Balance: towards a usable pervasive wellness application with accurate activity inference," in *Proceedings of the 10th workshop on Mobile Computing Systems and Applications*. ACM, 2009, p. 5.

[45] N. Oliver and F. Flores-Mangas, "Healthgear: Automatic sleep apnea detection and monitoring with a mobile phone." *JCM*, vol. 2, no. 2, pp. 1–9, 2007.

[46] Z. Jin, J. Oresko, S. Huang, and A. C. Cheng, "Hearttogo: a personalized medicine technology for cardiovascular disease prevention and detection," in *Life Science Systems and Applications Workshop, 2009. LiSSA 2009. IEEE/NIH*. IEEE, 2009, pp. 80–83.

[47] O. Akinbode, O. Longe, and B. Amosa, "Mobile-phone based patient compliance system for chronic illness care in nigeria," *Journal of Computer Science & Technology*, vol. 12, 2012.

[48] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 351–360.

[49] N. Bui, A. Nguyen, P. Nguyen, H. Truong, A. Ashok, T. Dinh, R. Deterding, and T. Vu, "Pho2: Smartphone based blood oxygen level measurement systems using near-ir and red wave-guided light," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 2017, p. 26.

[50] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data." in *Wireless Health*, 2016, pp. 30–37.

[51] C. Gao, F. Kong, and J. Tan, "Healthaware: Tackling obesity with health aware smart phone systems," in *Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on*. Ieee, 2009, pp. 1549–1554.

[52] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 341–350.

[53] T. Hao, G. Xing, and G. Zhou, "isleep: unobtrusive sleep quality monitoring using smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2013, p. 4.

[54] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 375–384.

[55] O. Amft, H. Junker, and G. Troster, "Detection of eating and drinking arm gestures using inertial body-worn sensors," in *Wearable computers, 2005. proceedings. ninth ieee international symposium on*. IEEE, 2005, pp. 160–163.

[56] S. Nirjon, R. F. Dickerson, Q. Li, P. Asare, J. A. Stankovic, D. Hong, B. Zhang, X. Jiang, G. Shen, and F. Zhao, "Musicalheart: A hearty way of listening to music," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 2012, pp. 43–56.

[57] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2017, pp. 278–291.