ESPIRO: Natural Pulmonary Function Monitoring via Earphone-Acquired Speech

Yetong Cao[1] Dong Ma[2] Wentao Xie[3] Qian Zhang[3] Jun Luo[1]
[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]School of Computing and Information Systems, Singapore Management University, Singapore
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China
Email: yetong.cao@sdu.edu.cn, dongma@smu.edu.sg, {wentaox,qianzh}@cse.ust.hk, junluo@ntu.edu.sg

ABSTRACT

As a crucial tool for assessing health, *spirometry* provides valuable insights into pulmonary functions. Recent advancements have enabled more convenient measurements by shifting spirometry solutions from cumbersome clinical devices to portable devices. However, the forced maneuvers and burdensome procedures, which necessitate repeated maximal forced breathing, often lead to dizziness and discomfort, rendering them unsuitable for vulnerable populations. In this paper, we present ESPIRO (Earphone-enabled Speech sPIROmetry) system to furnish user-friendly pulmonary function monitoring for diverse populations. Basically, ESPIRO records normal speech using microphone-embedded earphones and characterizes pulmonary function-related glottal flow during speech production. ESPIRO advances existing spirometry solutions in i) leveraging *phonetics* to associate pulmonary function with glottal flow in normal speech, thereby eliminating the need for forced breathing; ii) identifying effective speech features according to physiological basis, ensuring reliable spirometry measurements; and iii) effectively addressing ambient noise, making it suitable for various real-world settings. Extensive experiments with 38 subjects on 18 commodity earphones confirm that ESPIRO accurately estimates pulmonary function indices in practice.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; • **Applied computing → Health informatics**.

KEYWORDS

Spirometry, earable sensing, phonetics, speech analysis, glottal flow.

ACM MobiCom'25, November 4-8, 2025, Hong Kong, China

ACM ISBN 979-8-4007-1129-9/2025/11.
https://doi.org/10.1145/3680207.3723477

ACM Reference Format:
Y. Cao, D. Ma, W. Xie, Q. Zhang, and J. Luo. 2025. ESPIRO: Natural Pulmonary Function Monitoring via Earphone-Acquired Speech. In *The 31st Annual International Conference on Mobile Computing and Networking (ACM MobiCom'25), November 4-8, 2025, Hong Kong, China.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3680207.3723477

1 INTRODUCTION

Pulmonary diseases present a significant global public health challenge, with nearly 1 in 13 people affected by asthma [21], over 200 million individuals grappling with chronic obstructive pulmonary (lung) disease, and a staggering 488.9 million people suffering from respiratory infections [60]. To address this pressing issue, regular *pulmonary function* (PF) tests are essential for diagnosis, monitoring, and treatment assessment. Currently, the most widely utilized PF test tool is *spirometry*, which measures air volume and velocity expelled from the lungs. Despite its evolution from cumbersome clinical devices to portable and mobile options [2, 24, 69], the testing procedure remains burdensome and challenging: users must exert maximal effort to perform a deep inhalation followed by at least 6 seconds of forced exhalation, because normal or arbitrary breathing does not adequately reflect PF [45, 84]. This effort-dependent nature renders such testing unsuitable for the elderly, children, and pregnant women [59], often causing breathlessness or coughing [67]. Additionally, the necessity for test results to meet multiple criteria [28] often entails repeated attempts [6], resulting in extended measurement times (15-30 minutes), which often cause dizziness, shortness of breath, or even fainting [34, 56, 72].

To overcome the burdensome procedures of traditional spirometry, researchers have explored cough and wheeze sounds [55, 67], but their susceptibility to environmental factors undermines sample consistency and precision. Motivated by the fact that changes in voice quality can indicate underlying lung and airway issues [37], researchers have further investigated the potential of analyzing more stable vocal features for diagnosing abnormal pulmonary conditions [5, 38, 62, 75, 84]. However, the practicality of these systems is questionable: users must articulate specific content at maximum volume in quiet environments until breath

depletion, while maintaining a consistent distance from the recording microphone to ensure stable audio characteristics. These constraints hinder their deployment in many scenarios involving ambient noise and body movements, severely confining their practicality. Consequently, it is imperative to explore alternative spirometry schemes that avoid forced maneuvers and constrained recording conditions for wide adoption across diverse populations in real-world settings.

The ongoing rapid expansion and widespread adoption of earables have opened up a unique opportunity for developing novel health monitoring systems [7, 10, 13, 22], leveraging their proximity to the human body for non-invasive, continuous, and highly integrated health tracking [9, 80, 88]. By utilizing embedded microphones to capture signals from the wearer without disruption from varying acoustic propagation distances, earables have the potential to address the aforementioned challenges associated with PF monitoring. Indeed, the versatility of earables has already enabled successful monitoring of heartbeats [7, 10, 22] and respiration [32, 48]. In particular, EarSpiro [82] utilizes earables to act as spirometers using forced breathing. Nonetheless, the potential to leverage earables for assessing PF without forced maneuvers remains an open problem.

Inspired by the fact that variations in PF cause changes in flow patterns at the glottis (opening between the vocal cords) during speech production [33], we take a major step forward by combining the advantages of normal speech and ubiquitous microphone-embedded earphones for natural PF assessment. Towards this goal, we face several challenges: on the *theoretical* side, while the link between PF and glottal flow has been noted, the specifics of their interplay are not fully understood, making it essential to probe deeper into their relationship for effective system design. Besides, though initial studies [5, 75, 84] suggest maximal volume speech holds unique features for PF assessment, these features are not reliably present in normal speech. On the *practical* side, microphones (albeit embedded in earphones) are susceptible to ambient noise, compromising accurate pulmonary data acquisition. Existing noise reduction techniques opt for preserving semantic characteristics instead of critical PF details, highlighting the need for a new noise mitigation scheme. Moreover, our preliminary analysis indicates that PF details significantly overlap with certain noise components, making effective noise separation nearly impossible.

In this paper, we first introduce a phonetics model that characterizes interactions between the lungs and glottis, which validates the feasibility of assessing PF from glottal flow in normal speech. Based on this model, we introduce ESPIRO; a novel PF monitoring system that leverages microphones commonly embedded in earphones to estimate PF indices without forced maneuvers, making daily at-home monitoring more convenient and user-friendly. Specifically,

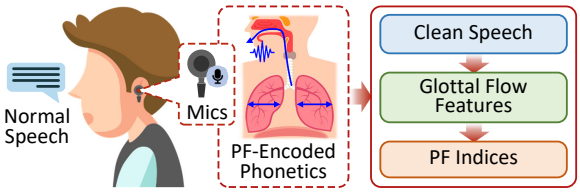


Figure 1: The concept of ESPIRO: inferring PF indices from glottal flow features in normal speech.

we first propose a novel scheme to identify speech segments that are qualified for PF indices estimation. We then develop a deep learning model with a loss function tailored for preserving critical PF information while eliminating ambient noise. Further scrutinizing the effects of PF variations on glottal flow patterns in the cleansed speech glottal flow enables us to establish effective features for capturing PF information. With these features, ESPIRO employs a deep neural estimator to achieve accurate PF indices estimation. To summarize, we make the following main contributions:

- We introduce ESPIRO, the first endeavour that interprets earable-captured *normal* speech into PF indices. ESPIRO can be readily integrated into microphone-embedded earphones; it offers a user-friendly alternative to existing methods requiring forced maneuvers.
- We associate glottal flow in normal speech to PF via a phonetics model. This model enables us to i) design novel algorithms to mitigate noise impact in normal speech while preserving critical PF details, and ii) establish effective glottal features for accurate PF estimation via a deep neural estimator.
- We conduct experiments with 38 subjects and 18 earphones. The results show that ESPIRO achieves errors within the ranges dictated by medical standards, suggesting its potential as a valuable tool for mobile pulmonary health management.

2 PRELIMINARIES

This section first introduces the background of spirometry and then explains how changes in PF can be reflected by glottal flow in normal speech.

2.1 Basics of Spirometry

Pulmonary diseases consistently alter the internal physiological conditions of the lungs and airways [90], leading to daily variations in physiological status, even in chronic cases [19], thereby necessitating frequent PF monitoring for prompt management and effective treatment. As a widely used test, spirometry indicates lung strength and breathing efficiency by measuring the volume and velocity of air an individual can exhale; it is crucial for diagnosing pulmonary diseases like asthma and chronic obstructive pulmonary disease (COPD).

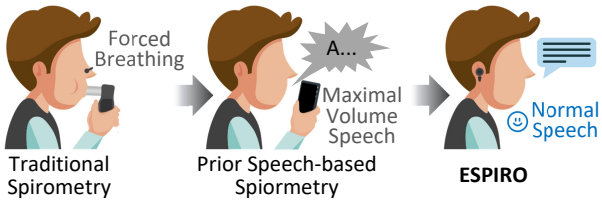


Figure 2: Spirometry shifts from forced breathing to maximal volume speech, and towards using ESPIRO.

Spirometry is typically conducted using a spirometer, a device with a mouthpiece linked to a computerized system to measure airflow volume and velocity. In practice, clinicians usually extract certain indices for more convenient disease evaluation and monitoring, including forced vital capacity (FVC), forced expiratory volume (FEV1), and FEV1/FVC.[1] These values are highly variable, influenced not only by pulmonary health status but also by factors such as age, height, and weight, with males generally exhibiting higher values than females [45]. Moreover, the test requires wearing a nose clip and performing maximal, forceful exhalations until the concerned breath is fully expended; failure to adhere to these instructions may result in inaccurate measurements. These forced maneuvers are challenging, particularly for vulnerable populations such as the elderly, children, and pregnant women. Additionally, the burdensome protocols result in high rejection rates of approximate 50% [6], extending the measurement duration to 15-30 minutes and often causing dizziness, shortness of breath, or even fainting [72].

Despite the evolution of spirometers from cumbersome clinical equipment to portable devices, forced maneuvers and burdensome procedures remain significant barriers to their widespread adoption. Therefore, this paper aims to overcome these barriers by utilizing earphone-captured normal speech to estimate PF indices, providing a user-friendly alternative to traditional spirometry, as shown in Figure 2.

2.2 From Speech to Pulmonary Function

Speech is one of the most common and essential activities in our daily lives. Despite variations in speaking speed and tone across different scenarios, normal speech generally exhibits stable patterns over months and years [42], which creates an opportunity for monitoring PF. Speech production involves the coordinated activity of lungs, vocal cords, and vocal tract (throat, oral cavity, nasal passages, and lips). We analyze their interactions to explain the rationale of ESPIRO.

As shown in Figure 3, the lungs first push airflow into the subglottal tract as a *lung pressure* wave p_I. After multiple reflections within the vocal tract, p_I transforms into p at the glottis, resulting in the *glottal flow* G (the specific correlation

[1]Peak expiratory flow (PEF) is not included in this study due to its significant variability and unreliability [54], as in other related works [69, 75, 84].

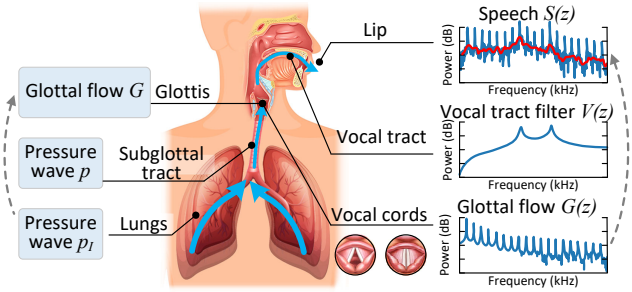


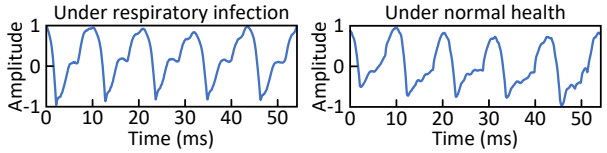
Figure 3: Speech production process.

between G, p, and p_I is explained in Section 3.2). The flow interacts with the vocal cords, leading to their periodic vibration and the production of the sound sources. These sound sources are subsequently filtered by the vocal tract, resulting in speech, a process that can be expressed mathematically in the z-domain [53]:

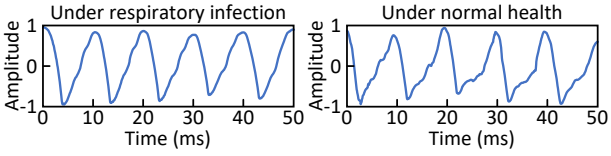
$$S(z) = G(z)V(z)L(z), \tag{1}$$

where $S(z)$ is the speech signal, $L(z)$ is the lip radiation, $V(z)$ is the vocal tract filter, and $G(z)$ is the glottal flow pulse. The above correlation, spanning from lung pressure wave p_I to glottal flow $G(z)$ and further to speech signal $S(z)$, reveals that *pulmonary information is encoded into normal speech via glottal flow—a relationship that previous related studies have not explored.* On this theoretical basis, we estimate PF function by analyzing the glottal flow in normal speech.

To validate the feasibility of this idea, we record normal speech from two subjects with lower respiratory tract infections, during illness and one week post-recovery (hence normal) to represent different pulmonary function states. Following the setup and pipeline in Sections 3 and 4, we process the speech recordings and obtain the time-domain waveform of glottal flow, which represents the glottal flow volume velocity. Figure 4 shows waveform examples for the phoneme /a:/ (as in 'Father'), which exhibits periodicity due to vocal cord vibration. Under the same pulmonary function conditions, each subject demonstrates stable patterns,



(a) Examples of glottal flow for subject 1.



(b) Examples of glottal flow for subject 2.

Figure 4: Glottal flow in different subjects: under respiratory infection vs. healthy state.

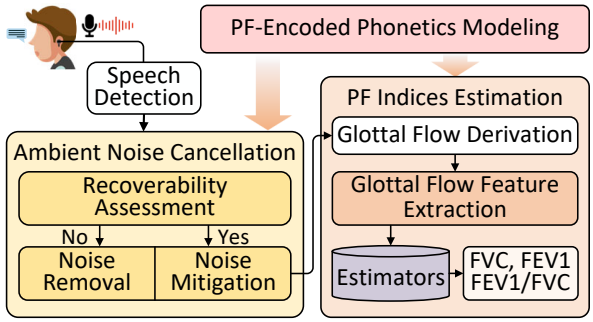


Figure 5: System architecture of ESPIRO.

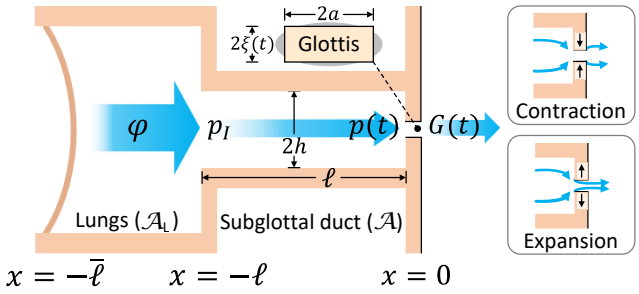


Figure 6: Mathematical model of the subglottal vocal region coupled to the glottis.

thereby validating the consistency of normal speech analysis. Comparing waveforms between under respiratory infection and normal health, we can observe significant differences related to peaks, troughs, and inflection points. This intra-class similarity and inter-class difference confirm the feasibility of distinguishing PF conditions exploiting glottal flow. Additionally, variability in glottal flow patterns among subjects highlights intricacies of this connection and the need for user-specific estimators.

3 ESPIRO: A MODEL-DRIVEN DESIGN

This section first provides an overview of ESPIRO's architecture and then delves into individual elements of its design.

3.1 System Overview

As illustrated in Figure 5, ESPIRO consists of four components: *PF-Encoded Phonetics Modeling*, *Speech Detection*, *Ambient Noise Cancellation*, and *PF Indices Estimation*. *PF-Encoded Phonetics Modeling* employs a widely recognized duct model to describe phonetics, which establishes a novel mapping from lung flow to glottal flow. According to this model, ESPIRO decodes PF indices from normal speech captured by microphone-embedded earphones: ESPIRO first detects human speech based on unique frequency cues in *Speech Detection*. Upon detecting speech, ESPIRO performs *Ambient Noise Cancellation* by classifying signals as recoverable or irrecoverable[2]; it then removes irrecoverable signals and mitigates noise in recoverable ones using a deep learning model with a loss function specifically designed to take PF information into account, ensuring reliable glottal flow characterization. Subsequently, in *PF Indices Estimation*, ESPIRO leverages an inverse filter-based method to derive glottal flow from the normal speech. It then extracts effective features by analyzing the response of glottal flow to PF variations using the phonetic model. Finally, a deep neural estimator is used to accurately estimate FVC, FEV1, and FEV1/FVC.

[2]Speech signals are deemed irrecoverable if noise and PF details are too heavily overlapped to allow for effective noise cancellation.

3.2 PF-Encoded Phonetics Modeling

As introduced in Section 2.2, PF information is encoded into speech via glottal flow. This is a relatively new and less studied area and still lacks sufficient theoretical justification. To bridge this gap, we study phonetics, which examines the interactions among vocal organs during speech. Specifically, we adopt a widely recognized duct model [30], as shown in Figure 6. The subglottal tract is modeled as a uniform, hard-walled duct of length ℓ and cross-sectional area $\mathcal{A}$ (width $2h$), extends from $-\ell$ to 0 on the x axis. At $x = -\ell$, it connects to an infinite duct representing the lungs, positioned at $x = -\bar{\ell}(< -\ell)$, $|\bar{\ell}| \to \infty$, with a cross-sectional area of $\mathcal{A}_L$. At $x = 0$, the subglottal tract connects to free space via the glottis, which features an elliptical shape but is modeled as rectangular with a constant span of $2a$ and a time-varying width of $2\xi(t)$. The following explores airflow dynamics through the lungs, subglottal duct, and glottis, along with their relationships.

PF determines lung flow. The steady contraction of the lung cavity produces a uniform volume flow in the form of a plane wave φ spanning the lung plenum, with constant strength q per unit area. It is determined by $\varphi = -c_0 q/2(t - x/c_0)H(t - x/c_0), -\bar{\ell} < x < -\ell$, with $H(t - x/c_0)$ denotes the Heaviside step function, c_0 is the sound speed. As an infinitely long duct, φ represents the lung's ability to deliver airflow—affected by lung elasticity, lung volume, and overall lung health—depends on q, which is a function of $\mathcal{A}_L$ [30]. This confirms that lung flow dynamics encode PF.

Lung flow establishes subglottal pressure waves. As the lung flow reaches the subglottal duct inlet, the rapid decrease in cross-sectional area reflects a delayed plane wave φ_R back into the lung plenum, which interacts with φ and generates a pressure wave $p_I = -\rho_0 \frac{\partial}{\partial t}(\varphi + \varphi_R) = \rho_0 c_0 q \mathcal{A}_L/(\mathcal{A}_L + \mathcal{A})$, where ρ_0 is the mean fluid density. This pressure wave then propagates through the subglottal duct, forming $p(t) = p_I H(t - x/c_0)$ at $x \sim 0$, corresponding to periodic expansion and contraction of the glottis. So far, the chain of correlations suggests that PF-reflected lung flow determines the pressure wave p_I and p in the subglottal duct.

Subglottal pressure waves form glottal flow. Airflow in the subglottal duct is ejected due to the pressure wave $p(t)$, forming the glottal flow at speed $v_\sigma(t)$ with the jet contraction ratio σ. According to the vortex sound equation [31], the relationship between $p(t)$ and $v_\sigma(t)$ satisfies

$$\frac{p(t)}{\rho_0} = \frac{\mathcal{A}}{2\pi c_0|\mathbf{x}|}\frac{\partial}{\partial t}\{\frac{2p_I}{\rho_0}\sum_{n=0}^{\infty}|\mathrm{R}|^n H(t-t_d) - \frac{1}{2}v_\sigma^2(t-t_e) - \sum_{n=0}^{\infty}\frac{1}{2}|R|^n[v_\sigma^2(t-t_d) + v_\sigma^2(t-t_d+\frac{2l_e}{c_0})]\}, \tag{2}$$

$$R = -\frac{\mathcal{A}_\mathrm{L}-\mathcal{A}}{\mathcal{A}_\mathrm{L}+\mathcal{A}}, t_d = \frac{[|\mathbf{x}|+2nL+\ell_e]}{c_0}, t_e = \frac{[|\mathbf{x}|+\ell_e]}{c_0}, |\mathbf{x}|\to\infty,$$

where the subglottal length is augmented by an end correction $L = \ell + \ell_e$ to better represent real conditions. Building on the earlier analysis, PF dictates q, subsequently affecting $p(t)$ and $v_\sigma(t)$. For homogeneous flow, the glottal volume velocity (the most commonly used time-domain representation for glottal flow [36]) is expressed as $G(t) = \sigma\Delta(t)v_\sigma$, where $\Delta(t) = 4a\xi(t)$ is the cross-sectional area downstream of the glottis. Therefore, by tracing these correlations, we can safely conclude that PF affects glottal flow, which can be quantified through the aforementioned equations. Clinical practice demonstrates that reduced PF, resulting in lung volume reduction and subglottal tube narrowing, typically leads to declines in FVC and FEV1 [91]—reductions in $\mathcal{A}_\mathrm{L}$ and $\mathcal{A}$ correspond to decreases in these indices. Furthermore, the FEV1/FVC ratio may either decrease or increase, depending on the specific disease [76]. With this theoretical basis, we design an effective speech processing pipeline to achieve accurate PF indices estimation.

3.3 Speech Detection

Accurate detection of speech is essential to avoid running expensive algorithms on non-speech periods. In real-world scenarios, the earphone's microphones can capture diverse ambient noise, making speech boundary detection very challenging. To address this, we leverage the fact that human speech demonstrates pronounced energy fluctuations across frequency bands, with distinct peaks at resonance frequencies, while non-speech audio signals demonstrate a more uniform energy distribution with gradual variations [95]. By analyzing energy changes between different frequency bands, we can accurately detect human speech. Specifically, we segment speech signals into 25 ms frames and calculate frame-wise sub-band variance. We then apply a dynamic threshold set at twice the amplitude variance of the previous three non-speech frames—an approach validated by prior research [95]—to accurately detect speech boundaries. In Figure 7, we showcase examples of clean speech, speech interfered by domestic and urban noises, and sub-band variance for three normal speech sentences. We can observe that frequency band variance peaks at speech boundaries and remains low in noise, and the dynamic threshold can robustly detect speech onsets and offsets, validating the effectiveness of our approach. Furthermore, to focus on normal speech, we exclude segments below 55 dB [15], thereby avoiding the processing of non-normal speech such as whispering.

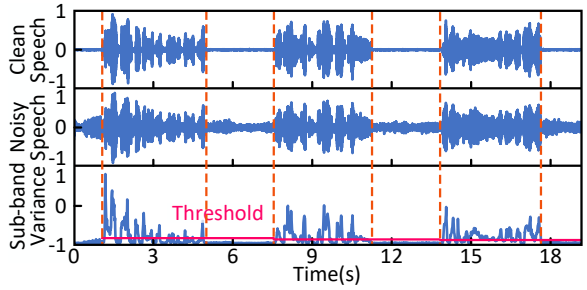


Figure 7: Illustration of speech detection and segmentation based on sub-band variance analysis.

3.4 Ambient Noise Cancellation

The obtained speech segments inevitably overlap with noise. To better serve the purpose of frequent PF monitoring, our goal is to mitigate the impact of noise while preserving PF details. However, certain noises—including nearby conversations, musical instruments, television/radio broadcasts, wind, car horns, and their combinations—can severely overlap with the speech spectrogram, making it impractical to remove noise without distorting the original speech. This is particularly challenging for glottal flow analysis, which is highly sensitive to spectral distortion, often leading to significant loss of PF detail [16]. Therefore, we introduce an innovative method for assessing signal recoverability, with specially designed reference signal selection to filter out signals with pronounced spectral overlap. Then, we apply deep neural networks to refine those recoverable signals for PF analysis.

3.4.1 Recoverability Assessment. Despite the availability of various speech quality assessment methodologies, their reliance on signal-to-noise ratio (SNR) or an overemphasis on intelligibility fails to adequately assess glottal flow distortions. Given the diversity of speech and noise, speech recoverability assessment remains a relatively novel and underresearched domain, lacking clear evaluation criteria. To address this challenge, we propose a novel comparator inspired by the success of pairwise comparisons of images and voices in scoring and ranking without clear standards [41, 44].

Comparator Structure. Our comparator, as shown in Figure 8, integrates three modules: embedding, temporal aggregation for extracting features, and binary classifier for comparing the recoverability of two signals. Specifically, i) the embedding model uses the Inception [70] network to

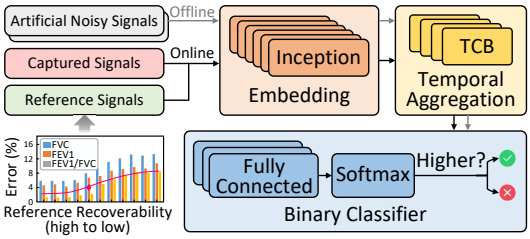


Figure 8: The architecture of comparator.

capture *multi-scale features* from the input signal, which effectively characterizes diverse normal speech without the need for selecting optimal kernel sizes. It employs six Inception blocks with parallel convolution layers of 1×1, 3×3, and 5×5 filter sizes, and an additional 3×3 max-pooling layer to extract more dominated features from the input. ii) The temporal aggregation model employs temporal convolutional networks to capture *long-term historical information* with low time and memory costs [35]. It comprises four temporal convolutional blocks (TCBs), each containing two convolutional layers with kernel size of 1×3 and channel sizes of 32, 64, 64, and 128, respectively. It also utilizes dilated convolutions with dilation factors of 2, 4, 8, and 16, along with weight normalization. iii) The binary classifier employs three fully connected layers and a softmax layer, classifying whether the recoverability of the second input signal is higher than the first input signal.

In the offline phase, we randomly select two clean speech signals from a public dataset [58] and introduce noise into them[3]. To control the level of spectral interference, we change factors affecting the noise spectrogram, such as altering the type of noise, adjusting noise power, filtering different frequency bands in the noise, or varying the mixing ratio between the noise and the speech signals, and ultimately obtain signals with different degrees of recoverability. These signals are then framed into 25 ms segments with 5 ms overlap and inputted into the comparator to facilitate training by minimizing the label-smoothed loss function:

$$L = \sum_{k=1}^{2} -(y_k(1-\alpha) + \tfrac{\alpha}{2})\log(p_k), \tag{3}$$

where p_k signifies the likelihood of classification into k_{th} class, while y_k denotes correct (1) or mis-classification (0) into respective class, with $\alpha = 0.1$ as an additional label smoothing parameter to prevent over-fitting [47]. In the online phase, earphone-captured speech segments are pairwise compared with reference signals that delineate recoverable and irrecoverable signals. Segments with recoverability lower than all references are classified as recoverable; otherwise, they are classified as irrecoverable.

[3]The noise is obtained from [51] and [58], encompassing animal sounds, natural soundscapes, human (non-speech) sounds, human speech sounds, domestic sounds, and urban noises.

Reference Signals Selection. Effective reference signals are critical for accurate recoverability assessment. Given that recoverable signals generally yield lower PF indices estimation errors compared to irrecoverable ones (as demonstrated in Section 4.3.2), we compare errors from various candidate signals to identify those that best delineate recoverable and irrecoverable signals. This process involves three iterative steps. i) *Establishing the Candidate Reference Signal Set*: We add noise to the R_{int} random clean speech signals, as we did earlier in the *offline* phase, to create signals with recoverable levels distributed from large to small. These signals are sequentially divided into g groups, from which a signal is selected to form a candidate reference set with different recoverable levels. ii) *Assessing Recoverability using Candidate Signals:* We collect PF indices and around 200 min diverse noisy normal speech signals from subjects. These noisy signals are compared with candidate signals using the comparator, and recoverable ones are used for PF indices estimation. iii) *Determination of the Optimal Reference Signal:* Errors trend in PF estimation reveals initially low errors at high recoverability levels of reference signals, indicating accurate classification of recoverable signals, although some may be missed. As recoverability in the reference signals decreases, errors rise sharply due to misclassification of irrecoverable signals as recoverable (see Figure 8). Clearly, the optimal reference signal is the one with the highest noise level before errors increase sharply, as it ensures accurate detection of most recoverable signals while minimizing the misclassification of irrecoverable ones. This iterative process is repeated R_{int}/g times to generate g reference signals with ample variability. In our proof-of-concept study, R_{int} and g are set to 120 and 10, but can be adjusted according to available computational resources. Section 4.3.1 validates the effectiveness of this method.

3.4.2 Noise Removal for Irrecoverable Signals. ESPIRO discards irrecoverable signals as it is impractical to remove noise interference from these signals while retaining PF detail. However, irrecoverable frames do not always occur continuously. If irrecoverable frames are interspersed with recoverable frames that are shorter than three frames, these recoverable frames are also removed as they lack sufficient contextual information for accurate pulmonary function analysis.

3.4.3 Noise Mitigation for Recoverable Signals. Despite advancements in speech noise mitigation [93], these methods are unsuitable for ESPIRO because their focus on semantics leads to spectrogram distortions, further introducing errors in glottal flow extraction. To address this issue, we build a deep model that uses a new deep feature loss [27] that compares crucial feature disparities in spectrograms between the denoised signal and a clean speech reference, thus avoiding distortions of PF information during noise mitigation.

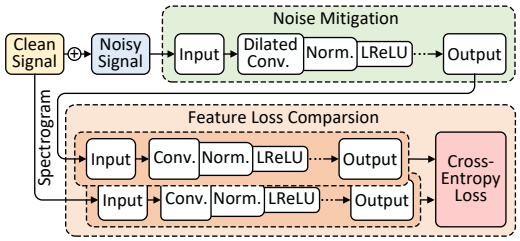


Figure 9: The architecture of noise elimination network based on deep feature loss.

Figure 9 shows the pipeline, which includes a noise mitigation component and a feature loss comparison component. They are trained alternately to minimize the spectrogram differences between real clean speech and the output signals.

Noise Mitigation Component. We design a cascaded convolutional architecture for noise mitigation, which demonstrates compactness and low runtime overhead in prior studies [12]. Specifically, the speech frames are fed into the input layer, then into 12 (experimentally determined) convolutional layers, utilizing 3×1 dilated kernel with dilation factors increasing from 2^0 in the 1-st layer to 2^{10} in the 11-st layer, and no dilation in the 12-nd layer. Each convolutional layer is followed by batch normalization (adaptive normalization) where a non-linear operation was implemented using point-wise non-linear leaky rectified linear units (LReLU) [12]. This architecture adapts to varying input signal lengths, enabling effective processing of speech sequences, including those with targeted portions located near sequence edges.

Deep Feature Loss Comparison. Using traditional loss functions like mean squared error (MSE) in noise mitigation fails to prevent spectrogram distortions, which hinders the subsequent glottal information extraction. To address this, we introduce a feature loss comparison component [27], imposing constraints on spectrogram feature differences between denoised and clean signals to guide training: It comprises 6 convolutional layers of kernel size 3×1, batch normalization, and LReLU units, with down-sampling between them by the factor of 2. During training, noise is added to clean speech signals s from open-source datasets, resulting in x. Then, x is processed by the noise mitigation network $\mathcal{N}$ with weight Φ to yield a quasi-clean speech signal $\mathcal{N}(x;\Phi)$. Subsequently, spectrograms of s and $\mathcal{N}(x;\Phi)$ are fed into the feature loss comparison module, resulting in cross-entropy loss as:

$$L_{s,x} = \sum_{m=1}^{6} \lambda_m \|\psi_m(s) - \psi_m(\mathcal{N}(x;\Phi))\|_1, \quad (4)$$

where ψ_m denote the m-th feature layer, with impact determined by λ_m set as the inverse of the relative values of $\|\psi_m(s) - \psi_m(g(x;\Phi))\|_1$ after 10 training epochs [27]. We showcase that deep feature loss produces more accurate spectrogram than MSE loss in Figure 10, affirming our approach's superiority. Section 4.3.2's experiments further underscore

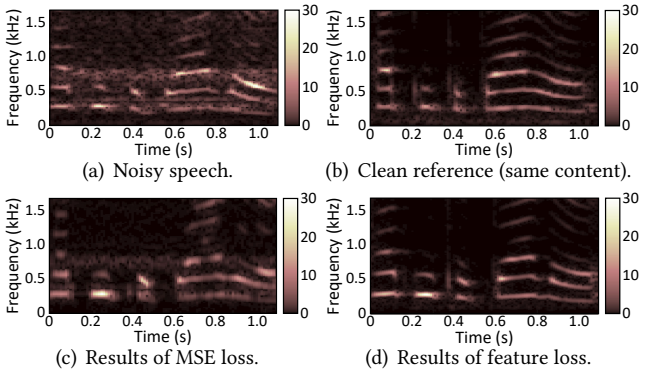


Figure 10: Feature loss efficiently mitigates noise, yielding more accurate speech spectrogram over conventional MSE loss. Low frequencies are shown to emphasize the noise reduction in the most affected range.

the improved performance of our method over conventional noise mitigation methods techniques.

3.5 PF Indices Estimation

Following the prior analysis, we analyze speech signals to extract glottal flow signals, then identify effective patterns by investigating the glottal flow variations under different PF conditions, and finally apply reliable regression techniques to predict PF indices.

3.5.1 Glottal Flow Derivation. Given the extensive research on glottal flow, we adopt a well-established Linear Predictive Coding (LPC)-based technique [3, 16], as shown in Figure 11. The employed multi-order LPC combinations and the setting of a 30 ms processing window have been extensively validated in the literature, thus eliminating the need for a redesigned extraction process. Specifically, we focus on phoneme /a:/ as an example (though other voiced phonemes are theoretically viable [84]), and we first detect phoneme /a:/ in clean speech via automatic speech recognition [36]. We follow these steps to extract glottal flow: i) apply 80 Hz high-pass filter on speech S to remove low-frequency interference unrelated to the speech and segment the filtered results into 25 ms frames. ii) perform first-order LPC on framed speech S to estimate Hg_1, representing the combined effects of the glottal flow and lip radiation. iii) inverse filter S with Hg and apply sixteenth-order LPC to roughly estimate vocal tract filter and obtain V', iv) based on V', inverse filter and integrate S to generate a preliminary glottal flow estimate, G'. v) repeat the above process with fourth-order LPC, inverse filtering, and integration, we accurately estimate the vocal tract filter V and obtain the final glottal flow estimate G by inverse filtering and integrating S.

3.5.2 Towards Effective Features. Based on the relationship between PF and glottal flow revealed by the phonetics model

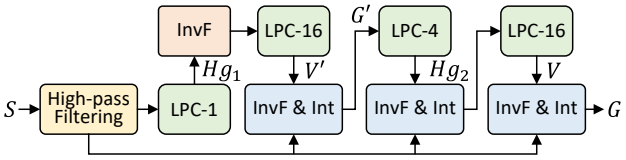


Figure 11: Extracting glottal flow from normal speech.

in Section 3.2, we proceed to delve into effective features for PF indices estimation, rather than directly feeding glottal flow into a neural network to learn a mapping that may not necessarily exhibit causality [63]. Applying Green's function [30] to solve Eqn. (2), we can derive the glottal volume velocity (time-domain glottal flow) $G(t)$ as:

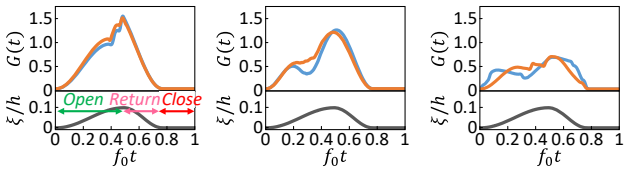
$$\frac{G(t)}{Q} = \sqrt{I^2 + \frac{2q\mathcal{A}_\mathrm{L}}{c_0\mathcal{A}} - \sum_{n=1}^{\infty}|R|^n[\frac{G(t-t_L)}{Q} + \frac{G(t-t_E)}{Q}]} - I, \quad (5)$$

$$Q = \sigma\Delta(t)c_0, I = \frac{Q}{c_0\mathcal{A}}, t_L = \frac{2nL}{c_0}, t_E = \frac{2nL - 2\ell_e}{c_0}, t \to \infty.$$

This equation clearly reveals that glottal flow depends on properties of lungs, subglottal duct, and glottis, including lung cross-sectional area $\mathcal{A}_\mathrm{L}$, subglottal duct cross-sectional area $\mathcal{A}$, and subglottal tube length L, with variations in amplitude associated with the time-varying glottal opening width $\xi(t)$. Among these factors, L is stable for a specific person, while $\mathcal{A}_\mathrm{L}$ and $\mathcal{A}$ are intrinsically linked to PF: Lungs' ability to deliver airflow and maintain exhalation (lung compliance) depends on $\mathcal{A}_\mathrm{L}$ [30]; $\mathcal{A}_\mathrm{L}$ and $\mathcal{A}$ also reflect lung capacity and the degree of obstruction [33]. Reductions in them indicate obstructive lung diseases (e.g., COPD and asthma), emphysema, restrictive lung diseases (e.g., pulmonary fibrosis), respiratory muscle weakness, increased airway resistance, or respiratory tract inflammation and infections [18, 29]. Hence, our focus shifts to exploring the response of $G(t)$ to $\mathcal{A}_\mathrm{L}$ and $\mathcal{A}$ dynamics as PF varies. Acknowledging that the iterative nature of Eqn. (5) convolutes the $\mathcal{A}_\mathrm{L}$-$\mathcal{A}$-$Q(t)$ relationship, we harness the junction reflection coefficient $|R| = |(\mathcal{A}_\mathrm{L} - \mathcal{A})/(\mathcal{A}_\mathrm{L} + \mathcal{A})|$ that involves both $\mathcal{A}_\mathrm{L}$ and $\mathcal{A}$ as an alternative means of exploration.

Using typical adult male parameters[4], we compute $G(t)$ based on Eqn. (5), as shown in Figure 12, where $|R|$ set to 0.9 and 0.5 to represent different PF states and $2f_0L/c_0$ set to 0.4, 0.6 and 0.8 to represent different subjects. During each glottal cycle, as glottis opens, returns, and closes, $G(t)$ increases to a peak value (representing the maximum flow rate when the glottis is fully open), decreases, and returns to zero, matching the waveform observed in real-world settings (Figure 4). Notably, different $|R|$ values cause significant differences in the trajectories of the glottal open and return

[4] p_I =10 cm of water (~ 1 kPa), f_0=125 Hz, c_0=340 m/s, ρ_0=1.23 kg/m^3, m = $0.5 \times 10^{-4}\,kg$, σ = 0.62 for $\mathrm{d}\xi/\mathrm{d}t > 0$, and σ = 1.15 for $\mathrm{d}\xi/\mathrm{d}t < 0$.



(a) $2f_0L/c_0$=0.4. (b) $2f_0L/c_0$=0.6. (c) $2f_0L/c_0$=0.8.

Figure 12: Normalized glottal volume velocity waveforms; $|R|$=0.9 (blue line), and $|R|$=0.5 (orange line).

phases, such as the number, positions, and amplitudes of peaks, troughs, and inflection points. Furthermore, the inherent variability in the properties, such as subglottal tract length, glottis width, and span, leads to different waveform patterns among subjects even under the same R, highlighting the need for user-specific PF estimators.

To effectively capture variations in glottal flow waveform, ESPIRO extract features from each glottal cycle, including *statistical features (mean, standard deviation, energy) of the number and amplitude of peaks, troughs, and inflection points, as well as the distance between them; media-crossing rate; and wavelet features (maximum, minimum, mean, and standard deviation of wavelet coefficients in each sub-band) from a four-level decomposition using db2.* These features are highly sensitive to waveform variations and are also widely used in existing studies on waveform analysis [11, 64]. We extract these features from the real-world data presented in Section 2.2, which covers two subjects experiencing healthy and impaired PF. We illustrate their t-SNE (t-distributed stochastic neighbor embedding) in Figure 13. Points from the same user exhibit distinct clustering under different lung functions, demonstrating that the features effectively differentiate between varying lung functions. Additionally, points from different users display unique clustering patterns, highlighting the necessity for developing user-specific estimators for each user.

3.5.3 Pulmonary Function Indices Estimation. After extracting effective features from glottal airflow, a new question emerges: *how should we estimate pulmonary function indices?* To answer this, we implement five well-established estimators from related work: support vector machine (SVM) [57], random forest regression (RFR) [62, 65], gradient boosting regression (GBR) [25, 62], adaboost regression (ABR) [62, 68], and convolutional neural networks with long short-term memory networks (CNN-LSTM) [75], each with different advantages and limitations. The first four models adopt empirically validate optimal configurations, and the CNN-LSTM model, consisting of cascaded 1D conventional layer, a dense layer, an attention layer, a concatenate layer, two LSTM layers, a dense layer, and finally a softmax layer, has been

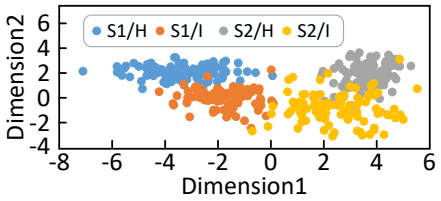


Figure 13: t-SNE projection of the features; H and I denote healthy and impaired PF, respectively.

demonstrated to effectively integrate CNN's prowess in nuanced feature analysis with LSTM's advantage in capturing long-term temporal dependencies [79]. Notably, we develop user-specific estimators that use personalized data from the users to infer FVC, FEV1, and FEV1/FVC (averaged from direct predictions and the ratio of predicted FEV1 to FVC). Cross-validation demonstrates that, despite performance variations, all indices fall within acceptable error ranges [69, 82]. We choose the CNN-LSTM model as the default estimator for its superior performance, as presented in Section 4.2.

4 EVALUATION

In this section, we first explain the implementation of ESPIRO along with the experiment setup. Then we conduct a thorough evaluation on ESPIRO under various scenarios.

4.1 Prototyping and Experiment Setup

Hardware and Software. ESPIRO employs microphone-embedded earphones as the sensing device to record normal speech sounds of the wearer. Specifically, the evaluation involves 18 pairs of earphones, each with varying prices around $100, featuring diverse microphone hardware and placements, thereby leading to variations in speech signal quality. Additionally, a laptop (with an Intel Core i7-12700H processor, 16 GB RAM, and an Nvidia GeForce RTX 4060 graphics card) processes the data as an end server to estimate pulmonary function indices: FVC, FEV1, and FEV1/FVC.

Data Collection. We recruit 38 participants, including 22 males and 16 females aging between 22 to 56 to collect normal speech data. Of these, 32 participants are healthy, while 6 have developed respiratory conditions during the study, providing a broad range of pulmonary function indices values. The study, approved by our university's ethics committee, uses a clinical-grade spirometer [17] to obtain baseline values for FVC, FEV1, and FEV1/FVC, following standard procedures [46]. Participants wear one of the available microphone-embedded earphones to record diverse normal speech, including both spontaneous and prepared readings (a subset of the LibriTTS corpus [94], which contains 50 short common English conversations featuring the vowel /a:/). They complete 6 to 10 sessions, each lasting 5 minutes each, resulting in a total of over 23,800 instances of /a:/ segments in the collected speech. We also record normal speech in various ambient noise conditions (30 to 70 dB), involving human (non-speech) sounds, human speech, domestic sounds, and urban noises, to assess the system's robustness to environmental noise and the effectiveness of the proposed noise mitigation technique. Additionally, we conduct a number of experiments across different scenarios and emotional states to validate ESPIRO's robustness and effectiveness.

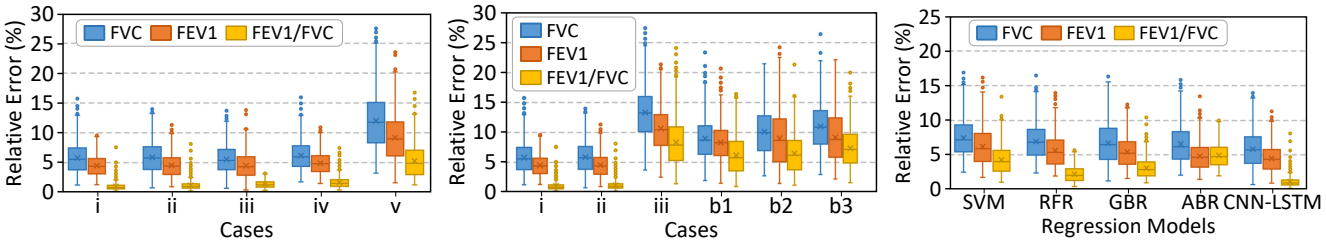
Table 1: Errors across different gender and age groups.

Gender	Index(%)	20-30	30-40	40-50	>50
Male	FVC	5.6±2.7	5.5±2.7	5.7±2.9	6.0± 2.8
	FEV1	4.0±2.0	4.3±2.1	4.7±2.1	4.8± 2.3
	FEV1/FVC	1.0±0.8	1.0±0.7	1.1±0.9	1.6±1.4
Female	FVC	5.6±2.9	5.7±3.1	6.3±2.9	6.4± 3.7
	FEV1	4.3±2.0	4.4±2.4	5.2±2.6	5.1± 2.6
	FEV1/FVC	1.0±1.2	1.3±1.8	1.4±1.6	1.8± 1.9

Evaluation Metrics. We assess the accuracy of estimating pulmonary function indices using the relative error between the estimated values x and the true values $\hat{x}$, which is defined as $(x-\hat{x})/\hat{x} \times 100\%$. For spirometry systems, the error ranges for FVC and FEV1 should be less than 10% [20, 69], while the error rang for FEV1/FVC is stricter, at below 5%, due to its composite nature [69, 82]. These serve as baselines to evaluate ESPIRO's performance.

4.2 Overall Accuracy

We build user-specific estimators for each subject based on their data from 18 different earphones and conduct five-fold cross-validation to evaluate ESPIRO's accuracy in inferring PF indices. The collected normal speech signals are processed as described in Section 3.4.3, which estimates FVC, FEV1, and FEV1/FVC, despite discarding roughly 12% of the instances due to noise reduction-related irrecoverability, with over 20,000 samples still remaining for the experiment. Overall, ESPIRO achieves average prediction errors of 5.8%, 4.6%, and 1.2% for FVC, FEV1, and FEV1/FVC, respectively. Since ESPIRO aims to serve diverse populations, Table 1 offers detailed averages±standard deviations, with results grouped by gender and age. We can observe distinct performance variations: errors are generally higher for females than males, and lower for middle-aged individuals than the elderly. Specifically, FVC errors range from 5.5% to 6.4% and FEV1 errors from 4% to 5.2%, both within the acceptable 10% error margin. Besides, FEV1/FVC exhibits lower errors due to its composite nature, ranging from 1% to 1.8%, within the acceptable 5% error margin. These results confirm ESPIRO's effectiveness in estimating pulmonary function indices across diverse populations.



(a) Recoverability Assessment. (b) Noise mitigation. (c) Features Design.

Figure 14: Individual components of ESPIRO contribute to reliable estimation of pulmonary function indices.

4.3 Effectiveness of Key Components

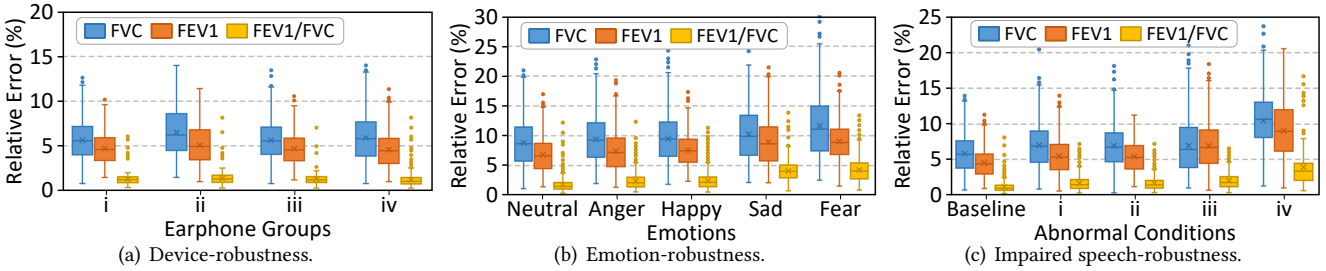
We study the effectiveness of key components in canceling ambient noise and estimating PF indices.

4.3.1 Effectiveness of Recoverability Assessment. Due to the lack of ground truth for recoverability, we validate our method by comparing pulmonary function indices estimation errors across datasets with varying degrees of recoverability. In particular, we use five datasets: i) clean speech signals; ii) signals classified as recoverable; iii) signals classified as recoverable with 10% randomly discarded; iv) signals classified as recoverable with 10% irrecoverable signals added; v) signals classified as irrecoverable. All datasets undergo noise elimination, and the resulting errors are illustrated in Figure 14(a). As expected, irrecoverable signals (case v) result in unacceptably high errors even after noise elimination, highlighting the need for precise recoverability assessment. Recoverable signals (case ii) show errors similar to clean signals (case i), indicating effective recovery to near-clean speech. Reducing the recoverable dataset (case iii) does not enhance performance, while increasing it with irrecoverable noise (case iv) raises errors. This validates the optimality of the current dataset and the effectiveness of the recoverability assessment. Furthermore, our analysis reveals that human speech sounds and urban noises in the surroundings typically result in irrecoverable signals, whereas human (non-speech) sounds and domestic sounds generally become irrecoverable only at higher noise levels.

4.3.2 Effectiveness of Noise Mitigation. Upon detecting recoverable signals, noise interference is mitigated for pulmonary function indices estimation. We particularly conduct five-fold cross-validation to evaluate ESPIRO's performance on (i) clean speech signals (about 6,000 s speech), (ii) recoverable signals after applying noise mitigation (about 5,200 s speech), and (iii) recoverable signals without applying noise mitigation (about 5,200 s speech). Additionally, to highlight the advantages of ESPIRO, we implement three traditional noise elimination methods as baselines: (b1) our method with MSE loss, (b2) spectral subtraction [73], and (b3) wiener filter [50]. The resulting relative errors are shown in Figure 14(b). For noisy speech signals without noise mitigation (case iii), the average relative errors for FVC, FEV1, and FEV1/FVC are 13.3%, 10.7%, and 8.4%, respectively, far exceeding the acceptable range. After applying noise mitigation (case ii), the errors significantly drop to 5.9%, 4.5%, and 1.2%, respectively, all within the acceptable error range and comparable to the errors of clean signals (case i). These results validate the effectiveness of our method in ensuring reliable estimation of pulmonary function indices. Besides, comparing case ii with b1 shows that without applying the deep feature loss, errors increase to 8.9%, 8.3%, and 6.2%, respectively. This indicates that our feature loss is beneficial and crucial in achieving the desired error margins. Additionally, case ii's lower errors compared to b2 and b3 demonstrate the superiority of our method over traditional ones.

4.3.3 Effectiveness of Features. We evaluate the effectiveness of the features by implementing five well-established estimation models, including SVM, RFR, GBR, ABR, and CNN-LSTM, with input layers adjusted to match our feature array size. Figure 14(c) illustrates the relative errors of these five models under five-fold cross-validation, with average errors for FVC of 7.2%,7.2%,6.3%,6.4%, and 5.8%; for FEV1 of 6.0%, 5.9%, 5.1%, 4.5%, and 4.6%; for FEV1/FVC of 4.0%, 2.5%, 2.8%, 4.6%, and 1.2%, respectively, all within the acceptable error range. These good cross-model performances validate the effectiveness of the features used by ESPIRO, indicating a strong correlation with pulmonary function indices. Additionally, we carefully study the feature importance of each model and find that the statistical features of the number of peaks, and amplitude of inflection points contribute most significantly to accurate estimations. Due to its lowest observed error, CNN-LSTM is selected as the default model. However, downstream applications may opt for other models based on accuracy requirements and computational resources, enabling ESPIRO to be deployed more broadly.

Moreover, we extract maximal volume speech features (MFCC, Mel-spectrogram features, etc.) validated in previous studies [75, 84] for experiments on normal speech. Among six models, CNN-LSTM attains the lowest average relative errors of 18.6%, 18.2%, and 9.7% for FVC, FEV1, and FEV1/FVC.



(a) Device-robustness. (b) Emotion-robustness. (c) Impaired speech-robustness.

Figure 15: ESPIRO exhibits robustness across various devices, emotional states, and impaired speech conditions.

However, these results lag significantly behind our proposed features, which directly reflect PF information, bypassing vocal tract modulation.

4.4 System Robustness

4.4.1 Impact of Earphone Models. Microphones are commonly integrated into earphones for functions such as call handling and noise cancellation. However, their varying placements affect sound propagation and consequently alter speech quality. To assess the robustness of ESPIRO across different earphone models, we specifically ask each participant to collect normal speech using 18 different earphones and compare system accuracy based on microphone positions: (i) 3 devices with on-cable microphones (integrated into the volume control buttons around chest height); (ii) 11 devices with on-ear microphones (integrated directly into the ear cups or earbuds); (iii) 4 devices with on-flexible arm microphones (mounted on a boom arm extending near the mouth); and (iv) all 18 devices. We conduct five-fold cross-validation, with both the training and testing data from the same device, with results shown in Figure 15(a). Group iii exhibits the lowest errors, with average FVC, FEV1, and FEV1/FVC of 5.6%, 4.6%, and 1.2%, respectively, likely thanks to the closer proximity of the microphone to the mouth. In contrast, group ii demonstrates the highest errors, potentially due to the longer sound propagation distance and potential noise from cable friction. Nevertheless, the errors remain within acceptable ranges, confirming that various earphone models have significant potential for reliable pulmonary function monitoring. Nevertheless, the errors remain within acceptable ranges and are comparable to earlier forced-breathing-based solutions (as demonstrated in Sec. 4.2). Thus, ESPIRO, an effortless speech-based system, emerges as a robust and viable alternative to prior solutions.

4.4.2 Impact of Emotions. Emotional states can influence glottal airflow patterns in speech [89], potentially affecting ESPIRO's performance. To evaluate system robustness under different emotions, we ask each participant to record normal speech while experiencing neutral, anger, happy, sad, and fear, triggered by video clips, a widely used method of eliciting emotions [66, 81]. Particularly, we train the estimator with neutral speech data, and evaluate it using data from these five emotions through five-fold cross-validations. As shown in Figure 15(b), the errors are generally higher under anger, happy, sad, and fear compared to the neutral state, with fear inducing the largest errors, likely due to changes in phonation from pressed to breathy and increased glottal flow waveform asymmetry [77]. Despite these variations, FVC remains below 7.7%, FEV1 below 6.0%, and FEV1/FVC below 2.9%, all within acceptable error ranges. This confirms that, while emotions influence the system, ESPIRO can reliably estimate pulmonary function indices.

4.4.3 Impact of Impaired Speech. In practice, impaired speech is inevitable and may hinder ESPIRO's performance. We particularly evaluate ESPIRO using impaired speech signals (about 2,240 sec) to assess the robustness of ESPIRO. These signals encompass (i) unclear pronunciation, (ii) speech repetition (stuttering), (iii) sudden pauses, and (iv) abnormal intonation. The relative errors are illustrated in Figure 15(c), with errors from normal speech serving as the baseline. Impaired speech generally results in higher errors than the baseline. Particularly, unclear pronunciation results in the highest errors, where FVC at 10.4%, FEV1 at 9.0%, and FEV1/FVC at 4.0%. While FEV1/FVC and FEV1 errors stay within acceptable ranges, the FVC error slightly exceeds the 10% threshold but is close to it. Errors for other types of impaired speech are within acceptable limits. These results suggest that ESPIRO effectively handles impaired speech and provides reliable pulmonary function estimates in most cases.

4.5 Performance on Subjects of Pulmonary Function Impairments

As introduced in Section 4.1, our dataset includes speech of six participants who developed respiratory diseases for 3 to 7 days and showed varying pulmonary function changes during the one-month experiment. We now evaluate ESPIRO's performance with these participants to demonstrate its potential in inferring pulmonary function indices for patients with respiratory conditions. Specifically, the training data

Table 2: Comparisons of ESPIRO with typical related works on mobile pulmonary function monitoring.

Items	SpiroSonic [69]	SprioSmart [39]	EarSpiro [82]	SpeechSpiro [75]	Yadav et.al [84]	ESPIRO
Measurement	Chest wall motion	Breathing sound	Breathing sound	Speech sound	Speech sound	Speech sound
Theoretical explanation	✓	✓	✓	×	×	✓
Void of extra hardware	✓	×	×	✓	✓	✓
Noise-robustness	✓	×	✓	×	×	✓
Unforced maneuvers	×	×	×	×	×	✓
FVC error	<2.5%RE	5.2%RE	9.9%RE	11% NRMSE	0.67L(of 1.7-2.63L)	5.8%RE
FEV1 error	<2.5%RE	4.8%RE	7.8%RE	12% NRMSE	-	4.5%RE
FEV1/FVC error	<2.5%RE	4.0%RE	5.1%RE	13%NRMSE	-	1.3%RE

are taken in healthy states, while the testing data comes from the week of illness (denoted as 0 weeks), and one, two, and three weeks post-illness. Figure 16 shows the average relative errors for FVC, FEV1, and FEV1/FVC among the six participants. Results demonstrate the significant potential of ESPIRO in tracking varying pulmonary function indices, with average relative errors for FVC below 6.2%, FEV1 below 4.8%, and FEV1/FVC below 1.3%, all within acceptable ranges.

4.6 Unique System Advantages

Table 2 compares ESPIRO with other typical mobile pulmonary function monitoring solutions, in terms of measurement targets, theoretical explanation, extra hardware requirements, noise robustness, necessity for forced maneuvers, and the reported errors for monitoring pulmonary function indices. Prior solutions, such as [39, 69, 82], primarily analyze chest expansion and breathing sounds associated with forced breathing to estimate pulmonary function indices. However, forced breathing poses significant barriers for vulnerable populations and often requires multiple attempts for accurate results, which could cause dizziness and discomfort. Besides, studies [39] and [82] necessitate specialized mouthpieces for effective sound collection, adding extra costs.

Studies [75, 84] demonstrate the feasibility of estimating spirometry indices by analyzing speech sounds, freeing users from forced breathing. However, they do not fully eliminate forced maneuvers, as users must read specific text at maximum volume until breath exhaustion. Additionally, their practicality is limited. They infer pulmonary function indices from common acoustic features, but the lack of theoretical justification raises concerns about the causality of their mappings. Furthermore, the requirement for a quiet environment and a microphone-enabled device at a specific distance restricts their use in noisy settings and scenarios involving body movements.

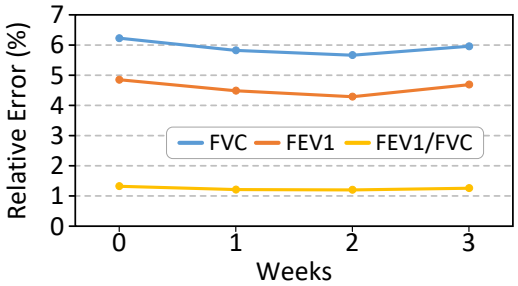


Figure 16: ESPIRO accurately tracks pulmonary function indices across various health conditions.

Compared to previous approaches, ESPIRO, a normal speech-based pulmonary function (PF) monitoring solution using earphones, offers several inherent advantages: i) ubiquitous deployment, leveraging commonly available microphone-embedded earphones and demonstrating robustness across different models; ii) user-friendliness, requiring only normal speech without forced breathing, thus eliminating the need for exhausting and forced maneuvers; iii) broad applicability, as prior systems, which depend on user effort, are less suitable for the elderly, children, pregnant women, and individuals with certain conditions, while normal speech is accessible to a wider population. Additionally, through rigorous theoretical research and system development, ESPIRO offers additional benefits, including a robust theoretical foundation that enhances medical reliability and efficient data processing algorithms that ensure effective operation in noisy environments.

4.7 System Overhead

To accurately assess system overhead, we deploy ESPIRO pipeline on a Raspberry Pi 4B platform, with DL algorithms converted to TensorFlow lite, simulating a standalone earable system with on-device processing. Particularly, the average system latency (looped 1,000 times) for recoverability assessment, noise mitigation, PF indices estimation, and other processes are 0.42s, 0.32s, 0.19s, and 0.27s, respectively. These results demonstrate that ESPIRO generates PF indices within 1.2s of user speech, validating its effective real-time performance. Besides, ESPIRO's speech recording energy consumption is about 181mw, and energy consumption of algorithms is 849mW (derived by subtracting the recording state power from the post-operational power). The microphone sampling runs continuously, but ESPIRO is only active for 1.2 s for

each estimate (despite around 12% of data experiencing extended delays due to noise interference), and an estimate is made around every 2.4 s (due to our focus on the prevalent phoneme /a:/). Consequently, the average energy consumed per second is 181 mW×1 s+849 mW×1.2 s/2.4=605 mJ.

5 DISCUSSION

As the first spirometry solution based on normal speech, ESPIRO certainly leaves several directions to be further explored. First, ESPIRO's noise robustness is limited by the use of a single microphone, making it effective only for processing recoverable signals with manageable interference. Given that advanced earphones now feature multiple microphones, it is prudent to explore their integration to enhance ESPIRO's noise robustness, potentially comparing it with future DL-based glottal flow refinement methods. Second, individual variability requires user-specific estimators, which may increase the burden of training data collection. We plan to address this by using labeled medical records and data augmentation techniques [74] to expand the dataset. Besides, we will employ continuous learning methodologies [8] and implement training data quality assessment to enhance the model's generalization capabilities and uphold training efficacy. Third, we are yet to deliver flow-volume curve [46], which some spirometers provide as a disease indicator [40]. However, since we already measure FVC and FEV1 values that correspond to specific points on the curve, it can be easily achieved through curve fitting [43]. Forth, due to IRB restrictions, we have not evaluated ESPIRO on subjects with severe pulmonary disease. Nonetheless, its strong performance in users with respiratory infections demonstrates its potential for tracking lung function variations. We are actively collaborating with hospitals to gather additional data from diagnosed patients. Fifth, ESPIRO currently focuses on the English phoneme /a:/. Studies [84] indicate potential for other phonemes, such as /i:/, /u:/, and /eɪ/, highlighting the need to explore PF inference across additional phonemes, languages, and dialects. Last but not least, in our pursuit of a user-friendly health monitoring system, gathering opinions from both users and medical experts is indispensable.

6 RELATED WORK

Clinical spirometry typically uses devices such as pneumotachs [23], ultrasonic sensors [78], and hot wire anemometers [52], which require expert operation and can be as large as a refrigerator [26]. Although portable spirometry devices [17] are available, they are either costly (> $2,000) or prone to high measurement errors (>20%).

Recent advances in mobile sensing offer new spirometry opportunities by using built-in sensors in mobile and wearable devices–such as cameras, microphones, and radio frequency sensors—to capture chest wall movements [4, 14, 24, 61, 69, 85–87, 92, 96], breath sounds [2, 39, 71, 82], and vibrations [1]. However, these systems require users to remain still, maintain an unobstructed chest view, and be in a noise-free environment, limiting their practical usability in real-life scenarios. Additionally, these solutions depend on forced breathing, which increases pressure in the chest, abdomen, and eyes, making them unsuitable for vulnerable individuals [59] and often necessitates multiple test iterations, potentially causing dizziness or shortness of breath [72].

To overcome the limitations of forced breathing, researchers have explored more natural approaches. Cough sounds [49, 59] and airway impulse response [91] have been successfully used for detecting pulmonary diseases. Additionally, speech analysis [38, 62] has been leveraged to infer pulmonary function indices. However, these methods still face significant limitations, such as requiring users to continuously say certain content until breath exhaustion, and their performance can be impaired by background noise in practical implementations. Compared to these existing solutions, ESPIRO avoids forced maneuvers and difficult procedures, offering advantages such as ubiquitous deployability, user-friendliness, applicability to diverse populations, and robustness against ambient noise. Notably, a parallel study has investigated PF monitoring via arbitrary breathing [83].

7 CONCLUSION

In this paper, we address the problems of forced maneuvers and difficult procedures associated with traditional spirometry. Particularly, we propose a novel pulmonary function monitoring system, ESPIRO, that enables frequent home-based monitoring by utilizing normal speech recorded by microphones-emended earphones to infer pulmonary function indices, including FVC, FEV1, and FEV1/FVC. By analyzing phonetics, ESPIRO reveals the implicit relationships between pulmonary function and glottal flow that generates speech. ESPIRO first filters the impacts of ambient noises in the captured normal speech, then extracts glottal flow that encodes pulmonary function information. After that, we propose effective features that accurately predict pulmonary function indices across various regression models. Real-world experiments confirm ESPIRO's accuracy and robustness, highlighting its potential to improve pulmonary function monitoring.

ACKNOWLEDGEMENT

We thank our anonymous shepherd and reviewers for their insightful comments and suggestions on this paper. This work was supported in part by the NSFC 62372045, MOE Tier 1 grant RG16/22, and RGC under Contract CERG 16206122, AoE/E-601/22-R.

REFERENCES

[1] Aakriti Adhikari, Austin Hetherington, and Sanjib Sur. 2021. mmFlow: Facilitating At-Home Spirometry with 5G Smart Devices. In *Proc. of the 18th IEEE SECON*. 1–9.

[2] Rishiraj Adhikary, Dhruvi Lodhavia, Chris Francis, Rohit Patil, Tanmay Srivastava, Prerna Khanna, Nipun Batra, Joseph Breda, Jacob Peplinski, and Shwetak Patel. 2023. SpiroMask: Measuring Lung Function Using Consumer-Grade Masks. 4, 1, Article 9 (feb 2023), 34 pages.

[3] Paavo Alku. 1992. Glottal Wave Analysis With Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication* 11, 2-3 (1992), 109–118.

[4] Edgar A Bernal, Lalit K Mestha, and Eribaweimon Shilla. 2014. Non Contact Monitoring of Respiratory Function via Depth Sensing. In *Proc. of IEEE BHI 2014*. 101–104.

[5] Sejal Bhalla, Salaar Liaqat, Robert Wu, Andrea S Gershon, Eyal de Lara, and Alex Mariakakis. 2023. PulmoListener: Continuous Acoustic Monitoring of Chronic Obstructive Pulmonary Disease in the Wild. *Proc. of ACM IMWUT 2023* 7, 3 (2023), 1–24.

[6] Brigitte M Borg, Moegamat Faizel Hartley, Mo T Fisher, and Bruce R Thompson. 2010. Spirometry Training Does Not Guarantee Valid Results. *Respiratory Care* 55, 6 (2010), 689–694.

[7] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones. In *Proc. of IEEE PerCom*. 200–209.

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

[9] Yetong Cao, Chao Cai, Fan Li, Zhe Chen, and Jun Luo. 2024. Enabling Passive User Authentication via Heart Sounds on In-Ear Microphones. *IEEE Transactions on Dependable and Secure Computing* (2024), 1–15.

[10] Yetong Cao, Chao Cai, Anbo Yu, Fan Li, and Jun Luo. 2023. EarACE: Empowering Versatile Acoustic Sensing via Earable Active Noise Cancellation Platform. *Proc. of ACM IMWUT 2023* 7, 2 (2023), 1–23.

[11] Yetong Cao, Qian Zhang, Fan Li, Song Yang, and Yu Wang. 2020. PPG-Pass: Nonintrusive and Secure Mobile Two-Factor Authentication via Wearables. In *Proc. of IEEE INFOCOM 2020*. 1917–1926.

[12] Qifeng Chen, Jia Xu, and Vladlen Koltun. 2017. Fast Image Processing With Fully-Convolutional Networks. In *Proc. of the IEEE International Conference on Computer Vision*. 2497–2506.

[13] Yiwei Chen, Wenhao Li, Xiuzhen Cheng, and Pengfei Hu. 2024. A Survey of Acoustic Eavesdropping Attacks: Principle, Methods, and Progress. *High-Confidence Computing* 4, 4 (2024), 100241.

[14] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust Vital Signs Waveform Recovery via Deep Interpreted RF Sensing. In *Proc. of the 27th ACM MobiCom*. 392–405.

[15] Decibel Meter Pro. 2024. Average Decibel Level of Human Speech. https://decibelpro.app/blog/how-many-decibels-does-a-human-speak-normally/ Accessed: 2024.

[16] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit. 2012. A Comparative Study of Glottal Source Estimation Techniques. *Computer Speech & Language* 26, 1 (2012), 20–34.

[17] e-LinkCare Meditech Co.,Ltd. 2021. UBREATH Spirometer System (PF680). https://www.e-linkcare.com/spirometer-system-pf680-product/ Accessed: 2024.

[18] JR Ederle, CP Heussel, J Hast, B Fischer, EJR Van Beek, S Ley, M Thelen, and HU Kauczor. 2003. Evaluation of Changes in Central Airway Dimensions, Lung Area and Mean Lung Density at Paired Inspiratory/Expiratory High-Resolution Computed Tomography. *European Radiology* 13 (2003), 2454–2461.

[19] Howard Eigen. 1999. Pulmonary function testing. In *Pediatric Asthma*. 131–150.

[20] P. Enright, W M Vollmer, B. Lamprecht, R. Jensen, and A S Buist. 2011. Quality of Spirometry Tests Performed by 9893 Adults in 14 Countries: The Bold Study. *Respiratory Medicine* 105, 10 (2011), 1507–1515.

[21] Asthma Facts and Figures. 2024. Key Facts about Asthma. https://www.aafa.org/asthma-facts/ Accessed: 2024.

[22] Xiaoran Fan, David Pearl, Richard Howard, Longfei Shangguan, and Trausti Thormundsson. 2023. APG: Audioplethysmography for Cardiac Monitoring in Hearables. In *Proc. of the 29th ACM Mobicom*. 1–15.

[23] Michael F Fitzpatrick, H McLean, AM Urton, A Tan, D O'donnell, and HS Driver. 2003. Effect of Nasal or Oral Breathing Route on Upper Airway Resistance During Sleep. *European Respiratory Journal* 22, 5 (2003), 827–832.

[24] Luay Fraiwan, Natheer Khasawneh, Khaldon Lweesy, Mennatalla Elbalki, Amna Almarzooqi, and Nada Abu Hamra. 2021. Non-contact Spirometry Using a Mobile Thermal Camera and AI Regression. *Sensors* 21, 22 (2021), 1–15.

[25] Jerome H Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* (2001), 1189–1232.

[26] Jake Garrison. 2018. *Spiro AI: Smartphone Based Pulmonary Function Testing*. Ph. D. Dissertation.

[27] Francois G Germain, Qifeng Chen, and Vladlen Koltun. 2018. Speech Denoising With Deep Feature Losses. In *Interspeech*. 1–6.

[28] Brian L Graham, Irene Steenbruggen, Martin R Miller, Igor Z Barjaktarevic, Brendan G Cooper, Graham L Hall, Teal S Hallstrand, David A Kaminsky, Kevin McCarthy, Meredith C McCormack, et al. 2019. Standardization of spirometry 2019 update. An official American thoracic society and European respiratory society technical statement. *American journal of respiratory and critical care medicine* 200, 8 (2019), e70–e88.

[29] V Hoffstein. 1986. Relationship Between Lung Volume, Maximal Expiratory Flow, Forced Expiratory Volume in One Second, and Tracheal Area in Normal Men and Women. *American Review of Respiratory Disease* 134, 5 (1986), 956–961.

[30] MS Howe and RS McGowan. 2009. Analysis of Flow-Structure Coupling in a Mechanical Model of the Vocal Folds and the Subglottal System. *Journal of Fluids and Structures* 25, 8 (2009), 1299–1317.

[31] Michael S Howe. 1998. *Acoustics of Fluid-structure Interactions*. Cambridge university press.

[32] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proc. of ACM IMWUT 2024* 8, 2 (2024), 1–25.

[33] Jenny Iwarsson, Monica Thomasson, and Johan Sundberg. 1998. Effects of Lung Volume on the Glottal Voice Source. *Journal of Voice* 12, 4 (1998), 424–433.

[34] Johns Hopkins Medicine. 2024. What Are the Possible Risks of Pulmonary Function Tests? https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/pulmonary-function-tests Accessed: 2024.

[35] Bee Hock David Koh, Chin Leng Peter Lim, Hasnae Rahimi, Wai Lok Woo, and Bin Gao. 2021. Deep Temporal Convolution Network for Time Series Classification. *Sensors* 21, 2 (2021), 603.

[36] Avinash Kumar, Syed Shahnawazuddin, and Gayadhar Pradhan. 2017. Improvements in the Detection of Vowel Onset and Offset Points in a Speech Sequence. *Circuits, Systems, and Signal Processing* 36 (2017), 2315–2340.

[37] S. Kumar and R. Salib. 2006. Upper Airway Obstruction. In *Encyclopedia of Respiratory Medicine*. Academic Press, Oxford, 375–385.

[38] John Kutor, Srinivasan Balapangu, Jeromy K Adofo, Albert Atsu Dellor, Christopher Nyakpo, and Godfred Akwetey Brown. 2019. Speech Signal Analysis as an Alternative to Spirometry in Asthma Diagnosis: Investigating the Linear and Polynomial Correlation Coefficient. *International Journal of Speech Technology* 22 (2019), 611–620.

[39] Eric C Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N Patel. 2012. SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone. In *Proc. of ACM UbiComp 2012*. 280–289.

[40] Jungsil Lee, Choon-Taek Lee, Jae Ho Lee, Young-Jae Cho, Jong Sun Park, Yeon-Mok Oh, Sang-Do Lee, and Ho Il Yoon. 2016. Graphic Analysis of Flow-Volume Curves: A Pilot Study. *BMC Pulmonary Medicine* 16, 1 (2016), 1–6.

[41] Jun-Tae Lee and Chang-Su Kim. 2019. Image Aesthetic Assessment Based on Pairwise Comparison a Unified Approach to Score Regression, Binary Classification, and Personalization. In *Proc. of the IEEE International Conference on Computer Vision*. 1191–1200.

[42] Anders Löfqvist and Bengt Mandersson. 1987. Long-Time Average Spectrum of Speech and Voice Analysis. *Folia Phoniatrica et Logopaedica* 39, 5 (1987), 221–229.

[43] WF Maddams. 1980. The Scope and Limitations of Curve Fitting. *Applied Spectroscopy* 34, 3 (1980), 245–267.

[44] Pranay Manocha, Zeyu Jin, and Adam Finkelstein. 2022. SQAPP: No-Reference Speech Quality Assessment via Pairwise Preference. In *Proc. of IEEE ICASSP 2022*. IEEE, 891–895.

[45] Martin R Miller, JATS Hankinson, Vito Brusasco, F Burgos, R Casaburi, A Coates, R Crapo, Pvd Enright, CPM Van Der Grinten, P Gustafsson, et al. 2005. Standardisation of Spirometry. *European Respiratory Journal* 26, 2 (2005), 319–338.

[46] James F Morris. 1976. Spirometry in the Evaluation of Pulmonary Function. *Western Journal of Medicine* 125, 2 (1976), 110–118.

[47] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When Does Label Smoothing Help? *Advances in Neural Information Processing Systems* 32 (2019), 1–10.

[48] Yunyoung Nam, Bersain A Reyes, and Ki H Chon. 2015. Estimation of Respiratory Rates Using the Built-in Microphone of a Smartphone or Headset. *IEEE journal of biomedical and health informatics* 20, 6 (2015), 1493–1501.

[49] Ebrahim Nemati, Md Juber Rahman, Erin Blackstock, Viswam Nathan, Md Mahbubur Rahman, Korosh Vatanparvar, and Jilong Kuang. 2020. Estimation of the Lung Function Using Acoustic Features of the Voluntary Cough. In *Proc. of the 42nd IEEE EMBC*. 4491–4497.

[50] Hilal H Nuha and Ahmad Abo Absa. 2022. Noise Reduction and Speech Enhancement Using Wiener Filter. In *International Conference on Data Science and Its Applications*. 177–180.

[51] Karol J Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proc. of the 23rd ACM International Conference on Multimedia*. 1015–1018.

[52] P Plakk, P Liik, and P-H Kingisepp. 1998. Hot-Wire Anemometer for Spirography. *Medical and Biological Engineering and Computing* 36, 1 (1998), 17–21.

[53] Michael David Plumpe, Thomas F Quatieri, and Douglas A Reynolds. 1999. Modeling of the Glottal Flow Derivative Waveform With Application to Speaker Identification. *IEEE Transactions on Speech and Audio Processing* 7, 5 (1999), 569–586.

[54] Santiago Quirce, Gustavo Contreras, Anne DyBuncio, and Moira Chan-Yeung. 1995. Peak Expiratory Flow Monitoring Is Not a Reliable Method for Establishing the Diagnosis of Occupational Asthma. *American Journal of Respiratory and Critical Care Medicine* 152, 3 (1995), 1100–1102.

[55] M V Achuth Rao, N K Kausthubha, Shivani Yadav, Dipanjan Gope, Uma Maheswari Krishnaswamy, and Prasanta Kumar Ghosh. 2017. Automatic Prediction of Spirometry Readings From Cough and Wheeze for Monitoring of Asthma Severity. In *2017 25th European Signal Processing Conference*. 41–45.

[56] Sharma Ravin. 2023. What Are The Side Effects Of Pulmonary Function Test? https://www.ganeshdiagnostic.com/blog/what-are-the-side-effects-of-pft Accessed: 2024.

[57] Jeffrey S Reynolds, W Travis Goldsmith, Jeremy B Day, Ayman A Abaza, Ahmed M Mahmoud, Ali A Afshari, Jacob B Barkley, E Lee Petsonk, Michael L Kashon, and David G Frazer. 2015. Classification of Voluntary Cough Airflow Patterns for Prediction of Abnormal Spirometry. *IEEE Journal of Biomedical and Health Informatics* 20, 3 (2015), 963–969.

[58] Carolyn Richie, Sarah Warburton, and Megan Carter. 2009. *Audiovisual Database of Spoken American English*. Linguistic Data Consortium.

[59] Gowrisree Rudraraju, ShubhaDeepti Palreddy, Baswaraj Mamidgi, Narayana Rao Sripada, Y Padma Sai, Naveen Kumar Vodnala, and Sai Praveen Haranath. 2020. Cough Sound Analysis and Objective Correlation With Spirometry and Clinical Diagnosis. *Informatics in Medicine Unlocked* 19 (2020), 1–11.

[60] Saeid Safiri, Ata Mahmoodpoor, Ali-Asghar Kolahi, Seyed Aria Nejadghaderi, Mark JM Sullman, Mohammad Ali Mansournia, Khalil Ansarin, Gary S Collins, Jay S Kaufman, and Morteza Abdollahi. 2023. Global Burden of Lower Respiratory Infections During the Last Three Decades. *Frontiers in Public Health* 10 (2023), 1–15.

[61] Hirokazu Sakamoto, Hiroki Takamoto, Takemi Matsui, Tetsuo Kirimoto, and Guanghao Sun. 2020. A Non-contact Spirometer With Time-of-Flight Sensor for Assessment of Pulmonary Function. In *Proc. of the 42nd IEEE EMBC*. 4114–4117.

[62] Nazir Saleheen, Tousif Ahmed, Md Mahbubur Rahman, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, Erin Blackstock, and Jilong Kuang. 2020. Lung Function Estimation From a Monosyllabic Voice Segment Captured Using Smartphones. In *Proc. of the 22nd ACM MobileHCI*. 1–11.

[63] Leonardo Seoane and David E. Taylor. 2003. Medical Errors in the Intensive Care Unit: Can We Find the Black Box Before the Patient Crashes? *Critical Care Medicine* 31, 10 (2003), 2553–2554.

[64] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. Face-Mic: Inferring Live Speech and Speaker Identity via Subtle Facial Dynamics Captured by AR/VR Motion Sensors. In *Proc. of the 27th ACM Mobicom*. 478–490.

[65] Paul F Smith, Siva Ganesh, and Ping Liu. 2013. A Comparison of Random Forest Regression and Multiple Linear Regression for Prediction in Neuroscience. *Journal of Neuroscience Methods* 220, 1 (2013), 85–91.

[66] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing* 3, 2 (2012), 211–223.

[67] Mateusz Soliński, Michał Łepek, and Łukasz Kołtowski. 2020. Automatic Cough Detection Based on Airflow Signals for Portable Spirometry System. *Informatics in Medicine Unlocked* 18 (2020), 100313.

[68] Dimitri P Solomatine and Durga L Shrestha. 2004. AdaBoost. RT: A Boosting Algorithm for Regression Problems. In *IEEE international joint conference on neural networks*, Vol. 2. 1163–1168.

[69] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: Monitoring Human Lung Function via Acoustic Sensing on Commodity Smartphones. In *Proc. of the 26th ACM MobiCom*. 1–14.

[70] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *Proc. of the IEEE CVPR*. 1–9.

[71] Sudipto Trivedy, Manish Goyal, Prasanta R Mohapatra, and Anirban Mukherjee. 2020. Design and Development of Smartphone-Enabled Spirometer With a Disease Classification System Using Convolutional Neural Network. *IEEE Transactions on Instrumentation and Measurement* 69, 9 (2020), 7125–7135.

[72] WT Ulmer. 2003. Lung Function–Clinical Importance, Problems, and New Results. *Journal of Physiology and Pharmacology: An Official Journal of the Polish Physiological Society* 54 (2003), 11–13.

[73] Navneet Upadhyay and Abhijit Karmakar. 2015. Speech Enhancement Using Spectral Subtraction-Type Algorithms: A Comparison and Simulation Study. *Procedia Computer Science* 54 (2015), 574–584.

[74] David A Van Dyk and Xiao-Li Meng. 2001. The Art of Data Augmentation. *Journal of Computational and Graphical Statistics* 10, 1 (2001), 1–50.

[75] Korosh Vatanparvar, Viswam Nathan, Ebrahim Nemati, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. SpeechSpiro: Lung Function Assessment From Speech Pattern as an Alternative to Spirometry for Mobile Health Tracking. In *Proc. of the 43rd IEEE EMBC*. 7237–7243.

[76] Carlos A Vaz Fragoso, John Concato, Gail McAvay, Peter H Van Ness, Carolyn L Rochester, H Klar Yaggi, and Thomas M Gill. 2010. The Ratio of FEV1 to Fvc as a Basis for Establishing Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* 181, 5 (2010), 446–451.

[77] Erkki Vilkman, Paavo Alku, and Juha Vintturi. 2002. Dynamic Extremes of Voice in the Light of Time Domain Parameters Extracted From the Amplitude Features of Glottal Flow and Its Derivative. *Folia Phoniatrica et Logopaedica* 54, 3 (2002), 144–157.

[78] Julia AE Walters, Richard WOOD-BAKER, Justin Walls, and David P Johns. 2006. Stability of the EasyOne Ultrasonic Spirometer for Use in General Practice. *Respirology* 11, 3 (2006), 306–310.

[79] Ke Wang, Changxi Ma, Yihuan Qiao, Xijin Lu, and Sheng Dong. 2021. A Hybrid Deep Learning Model With 1D CNN-LSTM-Attention Networks for Short-Term Traffic Flow Prediction. *Physica A: Statistical Mechanics and its Applications* 12 (2021), 1–13.

[80] Yong Wang, Tianyu Yang, Chunxiao Wang, Feng Li, Pengfei Hu, and Yiran Shen. 2024. BudsAuth: Towards Gesture-Wise Continuous User Authentication Through Earbuds Vibration Sensing. *IEEE Internet of Things Journal* (2024).

[81] Wanhui Wen, Guangyuan Liu, Nanpu Cheng, Jie Wei, Pengchao Shangguan, and Wenjin Huang. 2014. Emotion Recognition Based on Multi-Variant Correlation of Physiological Signals. *IEEE Transactions on Affective Computing* 5, 2 (2014), 126–140.

[82] Wentao Xie, Qingyong Hu, Jin Zhang, and Qian Zhang. 2023. EarSpiro: Earphone-based Spirometry for Lung Function Assessment. *Proc. of ACM IMWUT 2023* 6, 4 (2023), 1–27.

[83] Chi Xu, Wentao Xie, Baichen Yang, Yizhen Zhang, Yanbin Gong, Jin Zhang, Wei Li, Shifang Yang, and Qian Zhang. 2025. EasySpiro: Assessing Lung Function via Arbitrary Exhalations on Commodity Earphones. In *Proc. of the 31st ACM MobiCom*. 1–16.

[84] Shivani Yadav, Dipanjan Gope, Uma Maheswari Krishnaswamy, and Prasanta Kumar Ghosh. 2021. Convolutional Dense Neural Network Based Spirometry Variable FVC Prediction Using Sustained Phonations. In *Proc. of the 31st IEEE MLSP*. 1–6.

[85] Yanni Yang, Jiannong Cao, and Xiulong Liu. 2019. ER-rhythm: Coupling exercise and respiration rhythm using lightweight COTS RFID. *Proc. of the ACM IMWUT* 3, 4 (2019), 1–24.

[86] Yanni Yang, Jiannong Cao, Xiulong Liu, and Xuefeng Liu. 2019. Multi-Breath: Separate Respiration Monitoring for Multiple Persons With UWB Radar. In *Proc. of IEEE COMPSAC*, Vol. 1. IEEE, 840–849.

[87] Yanni Yang, Jiannong Cao, and Yanwen Wang. 2021. Robust RFID-based respiration monitoring in dynamic environments. *IEEE Transactions on Mobile Computing* 22, 3 (2021), 1717–1730.

[88] Yanni Yang, Pengfei Hu, Jiaxing Shen, Haiming Cheng, Zhenlin An, and Xiulong Liu. 2024. Privacy-Preserving Human Activity Sensing: A Survey. *High-Confidence Computing* (2024), 100204.

[89] Xiao Yao, Wensong Bai, Yuqian Ren, Xin Liu, and Zhijian Hui. 2020. Exploration of Glottal Characteristics and the Vocal Folds Behavior for the Speech Under Emotion. *Neurocomputing* 410 (2020), 328–341.

[90] Xiangyu Yin, Kai Huang, Erick Forno, Wei Chen, Heng Huang, and Wei Gao. 2023. PTEase: Objective Airway Examination for Pulmonary Telemedicine using Commodity Smartphones. In *Proc. of the 21st ACM MobiSys*. 110–123.

[91] Xiangyu Yin, Kai Huang, Erick Forno, Wei Chen, Heng Huang, and Wei Gao. 2023. PTEase: Objective Airway Examination for Pulmonary Telemedicine using Commodity Smartphones. In *Proc. of the 21st ACM Mobisys*. 110–123.

[92] Gu Yu, Meng Wang, Peng Zhao, Yantong Wang, Hao Zhou, Yusheng Ji, and Celimuge Wu. 2022. SpiroFi: Contactless Pulmonary Function Monitoring using WiFi Signal. In *Proc. of the 30th IEEE/ACM IWQoS*. 1–10.

[93] Asri Rizki Yuliani, M Faizal Amri, Endang Suryawati, Ade Ramdan, and Hilman Ferdinandus Pardede. 2021. Speech Enhancement Using Deep Learning Methods: A Review. *Jurnal Elektronika dan Telekomunikasi* 21, 1 (2021), 19–26.

[94] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A Corpus Derived From Librispeech for Text-To-Speech. *arXiv preprint arXiv:1904.02882* (2019).

[95] Xueying Zhang, Zhefeng Zhao, and Gaofeng Zhao. 2006. A Speech Endpoint Detection Method Based on Wavelet Coefficient Variance and Sub-Band Amplitude Variance. In *Proc. of IEEE ICICIC 2006*, Vol. 3. 83–86.

[96] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. 2021. MoRe-Fi: Motion-robust and Fine-grained Respiration Monitoring via Deep-Learning UWB Radar. In *Proc. of the 19th ACM SenSys*. 111–124.