EasySpiro: Assessing Lung Function via Arbitrary Exhalations on Commodity Earphones

Chi Xu[†*], Wentao Xie[†*], Baichen Yang[†], Yizhen Zhang[†], Yanbin Gong[†], Jin Zhang[‡], Wei Li[♯], Shifang Yang[§◇], Qian Zhang[†◇]

[†]The Hong Kong University of Science and Technology, [‡]Southern University of Science and Technology
[♯]Infectious Diseases Dept, Union Hospital, Tongji Medical College, Huazhong Univ of Science and Technology
[§]Guangdong Provincial People's Hospital and Southern Medical University

{cxubs, wentaox, byangak, yzhangtf, ygongae, qianzh}@cse.ust.hk,
zhangj4@sustech.edu.cn, lsnlw@sina.com, yangshifang@gdph.org.cn

Abstract

Conventional pulmonary function tests (PFTs) are important but costly. Hence, prior research has proposed IoT sensor-based solutions to facilitate cost-efficient, at-home PFT. However, these solutions require the subject to perform maximal exhalations, a task often challenging without supervision, compromising test accuracy. In response to this challenge, this study introduces EasySpiro that, for the first time, uses non-maximal exhalations to measure PFT indicators. This is challenging since PFT indicators are only defined for maximal exhalations, and there are no guidelines to derive them from submaximal exhalations. To address that, we observe that pulmonary deficiencies affect all types of breathing, where the underlying pulmonary deficiency should be the same under different breathing efforts. Leveraging this insight, we design a reconstruction model to predict the ideal maximal breathing patterns based on submaximal ones and utilize these reconstructions for PFT. Furthermore, since the body dynamics reflect the exhalation effort, we use self-supervised learning techniques to encode body dynamics into breathing effort representations to guide the reconstruction process. We integrate these designs into earphones with microphones to measure breathing patterns and IMUs to measure body dynamics. We collaborate with a hospital and develop a dataset from 50 patients with various diseases to evaluate EasySpiro's performance, which shows an accurate prediction of PFT indicators based on non-maximal exhalations with an error rate of 7%. In addition, we open-source the collected dataset to encourage future research.

[*] Co-first authors. [◇] Co-corresponding authors.

ACM MOBICOM '25, Hong Kong, China

ACM ISBN 979-8-4007-1129-9/2025/11
https://doi.org/10.1145/3680207.3723489

CCS Concepts

• **Human-centered computing** → **Mobile devices**; • **Computing methodologies** → *Machine learning*.

Keywords

Spirometry, breathing acoustics, signal reconstruction

ACM Reference Format:
C. Xu, W. Xie, B. Yang, Y. Zhang, Y. Gong, J. Zhang, W. Li, S. Yang, Q. Zhang. 2025. EasySpiro: Assessing Lung Function via Arbitrary Exhalations on Commodity Earphones. 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/10.1145/3680207.3723489

1 Introduction

The pulmonary function test (PFT) is the golden standard for measuring how well the lungs work and evaluating respiratory function. Conventional PFTs are conducted through spirometry, where the user is instructed to exhale with their maximal effort for at least six seconds into a medical device called a spirometer, which examines the airflow properties and outputs a set of lung function indicators to characterize a person's lung condition [18]. However, because of the high cost of spirometers (usually > 2,000 USD) and PFT diagnosis, not all users can afford to conduct PFT frequently.

Fortunately, previous works [1, 17, 25, 28, 52, 66] have significantly reduced the cost of a PFT by replacing spirometers with IoT sensors. Those works leverage sensing devices such as earphones and phones to capture the airflow properties during a forced exhalation maneuver to achieve similar functionality as a spirometer through audio or posture analysis. In this way, at-home PFTs can be enabled cost-efficiently.

Nevertheless, the above work only solved half of the problem - even if the users have low-cost PFT devices, they still can hardly perform valid PFTs at home. This is because the standard PFT requires maximal effort breathing maneuvers (Figure 1(a)), which users cannot perform alone and need clinicians to teach. Otherwise, low-quality PFT results may

(a) Traditional PFT (b) EasySpiro

Figure 1: Conducting PFT. (a) Traditional method (b) EasySpiro.

occur due to submaximal effort, making them unusable [18]. In fact, even in a clinical setup, it is essential to ensure the quality of the test since the rejection rate of PFT can be as high as 50% [5], due to the challenges in performing the test. Therefore, multiple PFT maneuvers are required to guarantee the usability of the test which is supervised by the clinician. Furthermore, there are some populations, such as the elderly, patients with severe heart disease and high blood pressure, and pregnant people, who are not suggested to perform standard PFT due to the risk that repeated PFT tests may cause shortness of breath and fainting, while their lung function also needs to be assessed properly [43, 55, 59].

In other words, it is quite challenging for users to perform the standard maximal-effort PFT at home alone. However, it is relatively easy for everyone to perform submaximal exhalation, just like casual deep breaths. From this perspective, we raise a question: *can we use the submaximal breathing patterns to infer pulmonary function?* A good message based on our observation is that although the standard PFT is based on maximal exhalation, medical research has shown that breathing with less effort, even normal breathing, will also exhibit different characteristics among different pulmonary conditions [4, 31, 40, 65]. Thus, extracting features from the breathing patterns of submaximal breathing efforts can potentially infer pulmonary functions. Inspired by EarSpiro [66], we use microphones embedded in earphones to collect exhalation sounds as a surrogate for breathing patterns and use the measured audio patterns to infer indicators of pulmonary function.

Although it sounds promising, achieving this goal is challenging. The biggest challenge is modeling the relationship between pulmonary function and submaximal exhalations. As discussed above, the pulmonary function indicators are only defined in the context of maximal exhalations [18], and there are no pulmonary function indication methods for submaximal ones. Our core idea to solve this challenge is that lung deficiencies manifest anomalies in both maximal and casual breathing patterns. Hence, we propose a reconstruction method to predict the ideal, maximal exhalation patterns from the sub-optimal ones and use this prediction to estimate lung function. In this way, the lung function estimation will be based on maximal efforts, and this is aligned with the medical standard. Therefore, we design a UNet-based [45] image-to-image generation model to reconstruct the maximal exhalation patterns and use the reconstructed ones for the subsequent lung function indicator prediction.

Notably, in order to reconstruct a consistent maximal exhalation pattern based on the submaximal ones, we need first to measure the exhalation efforts to let the model know the gaps between the submaximal and maximal maneuvers. However, there is no explicit definition of breathing effort or explicit label for it. Therefore, the second challenge faced by our design is measuring and encoding the breathing effort. To solve this challenge, we observe that different breathing efforts will result in different posture dynamics. That said, we can use earphone-embedded IMUs to characterize breathing effort, and we use the IMU characterizations to guide the reconstruction process. To solve the issue of the lack of explicit breathing effort labels, we propose to use a self-supervised contrastive learning scheme to encode the breathing effort. By training an encoder network that can compare two breathing efforts, we can encode breathing efforts properly. In addition, to increase the representation ability of our model, we build the breathing effort encoder based on an IMU-based activity recognition model, LIMU-BERT [69], which is pretrained by large IMU datasets.

One additional challenge is that the breathing patterns and lung conditions are significantly different among the patients and healthy populations. Therefore, to develop the above-mentioned deep-learning models, a fairly large amount of data from both the patient and healthy population is required. In this research, we collaborate with a hospital to collect a dataset containing the exhalation sounds of diverse breathing efforts from 50 subjects containing 36 patients with various diseases. We build and evaluate our system on this dataset.

With the above designs, we present EasySpiro, an earphone-based PFT solution based on non-maximal exhalations. The using scenario of EasySpiro is depicted in Figure 1(b). The earphones are equipped with a pair of microphones to collect exhalation sound and a pair of IMUs to measure posture dynamics. We build and evaluate our system on our dataset, and the evaluation result shows an average pulmonary function indicator error of 7%. We summarize the contributions of this work as follows:

- We propose EasySpiro, the first PFT solution to predict pulmonary functions based on casual, non-maximal breathing efforts, whereas originally, PFT could only be conducted with a maximal breathing effort.

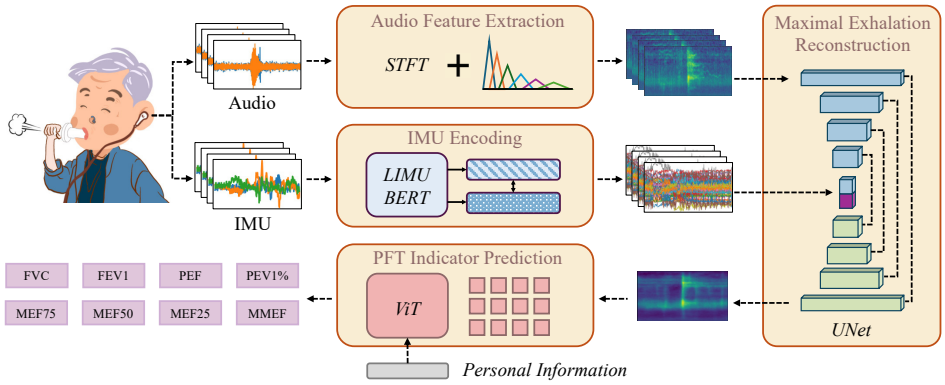


Figure 2: Overview of EasySpiro.

- We propose a series of techniques to solve a few challenges, including a spectrum reconstruction model to predict the ideal maximal exhalation patterns from the submaximal ones and a breathing effort encoding method based on contrastive learning and pretraining on large IMU datasets.
- We build EasySpiro on a pair of earphones and collaborate with a hospital to collect a dataset of patients with various diseases to validate our system.
- We open-source the first PFT dataset based on exhalation sounds. The dataset is available at the following link: https://github.com/ERICXUCHI/EasySpiroDataset.

2 Background

Before we dive into the details of our design, we first discuss the general background of this research. We begin with a brief introduction of PFTs. Then, we discuss the reason why we can use breathing patterns to infer pulmonary functions.

2.1 Pulmonary Function Test

PFTs are usually conducted through spirometry, where the subject exhales into a spirometer, which analyzes the airflow properties and outputs pulmonary function indicators. Note that here, the subjects are required to exert their maximum effort, as fast and hard as they can, to expose most of their lung restrictions [18]. The exhalation effort must be held for at least six seconds to make sure the air in the lungs is cleared. Because of the challenge of this maximal exhalation maneuver, the subjects often feel exhausted and even faint after the test. Therefore, a clinician must guide the subject to ensure the PFT quality. The primary outcome of the PFT is a set of pulmonary function indicators, such as PEF, FVC, FEV1, and FEV1/FVC, which are representative of different pulmonary conditions. Please refer to Graham *et al.* [18] for the definitions of these indicators. The target of this work is also to estimate these pulmonary function indicators.

2.2 From Casual Breathing Sound to PFT

Previous research has shown the feasibility of using breathing sounds as a surrogate for airflow speed [2, 17, 28, 66]. This is because when a subject breaths, turbulence will form in the constrictive portion of the airways, and this turbulence will generate sounds [27]. It has been demonstrated that the audio properties of the generated airflow sounds are correlated with the airflow speed [15]. Therefore, we can use the earphone-captured breathing sound to infer the breathing patterns so as to perform PFTs.

Although the standard PFT requires a maximal exhalation, recent medical research has shown that submaximal exhalations or even normal breathing can show different characteristics under different lung conditions [4, 31, 40, 65]. This is because, in patients with impaired lung functions, their airways are obstructed, which keeps air from moving in and out of the lungs freely. This will not only limit the maximal speed that one can exhale with, but also interrupt the normal breathing routine. For example, compared with a healthy subject, respiratory disease patients exhale slower even for normal breathing [35], and it takes a shorter time for the patients to reach the peak expiratory flow rate [34]. These works tried to extract consistent indicative metrics from casual breathing patterns as alternatives for PFTs. However, due to the heterogeneity of casual breathing and the limited study population, medical standards or guidelines are yet to be established. In this work, we wish to use the earphone-captured casual exhalation audios to conduct PFT, with the help of advanced machine learning and signal processing techniques.

3 System Design

This section elaborates on the details of EasySpiro. First, we will give an overview of our design. The system diagram is presented in Figure 2. The design comprises four main modules that process audio and IMU data cooperatively. In the

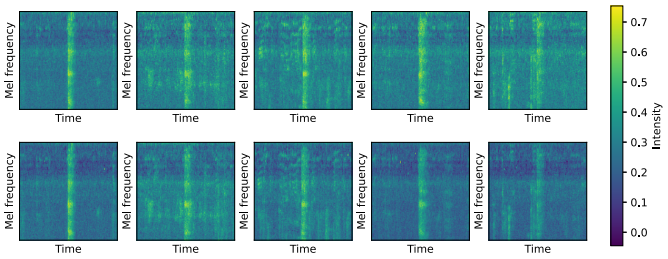


Figure 3: Multiple session spectrograms. (Upper row: left ear; Lower row: right ear)

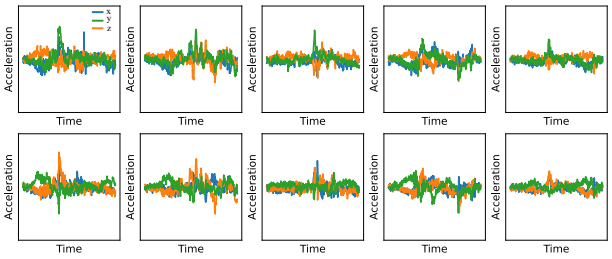


Figure 4: Multiple session IMU signals. (Upper row: left ear; Lower row: right ear. Colors represent axes.)

first module (Section 3.1), we extract spectral features from breathing sounds as the foundation of the PFT prediction using signal-processing techniques. Then, as discussed in Section 1, we use a reconstruction methodology to predict the ideal maximal exhalations to enable PFT that is aligned with medical guidelines. In this reconstruction process, we use IMU data to encode the information on the exhalation effort to guide the reconstruction. The rationale behind this design is that different exhalation efforts result in different body dynamics. Encoding this information in the loop hints at the reconstruction model on the gaps between the submaximal exhalation and the maximal one. Therefore, we design the second module (Section 3.2) to encode exhalation effort information from the IMU data through self-supervised learning techniques. After that, our third module (Section 3.3) leverages an image transformation model to reconstruct the maximal exhalation spectrogram from several arbitrary-effort ones. In the final module (Section 3.4), the reconstructed spectrogram is regarded as an image sample and processed by a state-of-the-art image processing model for PFT indicator prediction. The details of these modules are discussed in the subsequent sections.

3.1 Audio Feature Extraction

This module aims to extract sufficient features from breathing sounds to support subsequent modeling. After receiving the raw breathing audio, the first step is to locate the starting point of exhalation. A straightforward method to do so is to find the loudest part of the audio segment, as a PFT begins with a strong exhalation. However, this cannot work in some cases with a noisy environment. Fortunately, since we also attach the IMU sensor to the earphone, the IMU signal shows significant changes at the starting point of exhalation. Therefore, we can determine the starting point by analyzing the amplitude variation in the IMU signal.

Once we identify the starting point, we apply the Short-Time Fourier Transform (STFT) to extract time-frequency features using a Hanning window. However, we face the same issue of high-dimensional spectrograms as noted in EarSpiro [66]. To address this, we utilize a Mel Filter Bank to reduce the dimensionality [49, 74]. To emphasize the low-frequency components, which contain more informative content than high-frequency parts, we use a Mel filter bank with 64 filters to process the spectrogram. The parameters for the STFT and Mel filter bank are adapted from [73]. Specifically, we extract 12 seconds of raw audio starting from the exhalation points, using a window size of 64 ms and an overlap of 32 ms between consecutive segments.

Moreover, we observe that due to the low-cost hardware, the collected breathing sounds are often mixed with the hardware's internal noises. To address this, we record the ambient signals in a completely silent environment as a template and reduce the future recording's hardware noise by subtracting this template.

We show five sessions of the processed spectrograms in Figure 3, with effort decreasing from left to right. As discussed in Section 1, we will reconstruct the ideal, maximal-effort exhalation patterns based on these submaximal patterns. However, given their significant heterogeneity because of the arbitrary breathing efforts, it is hard for the system to give a consistent prediction. In the next two sections, we introduce our approach to achieving a high-quality reconstruction with the help of IMU data.

3.2 Exhalation Effort Encoding with IMU

As discussed previously, we use IMU signals to measure and encode the exhalation effort, which guides the reconstruction process. The basic principle of this idea is that the IMU readings will display a certain pattern capturing the user's level of effort. This pattern is characterized by significant variations in the IMU signal if the user swings a lot during PFT - a symbol of exhale forcefully. This happening means that the audio spectrograms that are associated with this IMU signal should closely resemble that of a maximal effort. Conversely, a lack of variation suggests a divergence from maximal effort. This understanding aids the model in identifying the generation gap effectively.

The core challenge, however, is how to encode such patterns that represent exhalation effort since there is no explicit label for it. Fortunately, self-supervised learning has shown

significant promise in handling non-annotated datasets [22]. Therefore, we use self-supervised learning techniques to encode the exhalation effort. The rationale behind this design is that although no labels for exhalation effort are available, we know which exhalation effort is greater between two exhalation maneuvers. This way, we can encode the exhalation effort by training an IMU encoder using contrastive learning that compares the effort level between two exhalation maneuvers. We present five sessions of IMU signals from one person in Figure 4, with effort decreasing from left to right. We can indicate from this figure that more exhalation effort results in a larger variation in the IMU data, and this IMU data can serve as the measurement of the gap of the reconstruction.

Based on these assumptions, we adopt two self-supervised learning schemes to process the IMU data. First, we use a state-of-the-art IMU processing model, LIMU-BERT [69], as the backbone of our encoder. We pretrain this model on large IMU datasets [32, 44, 50, 53] using random masking as recommended in [69]. Second, to encode exhalation effort from IMU data, we design a contrastive learning scheme. Specifically, we randomly select two IMU signals from one subject, along with their corresponding peak expiratory flow (PEF) values, since medical research has shown that the PEFs can be used to compare exhalation effort. The positive and negative samples in the contrastive learning scheme are therefore defined in Equation 1.

$$label = \begin{cases} 0 & : PEF_{IMU_1} > PEF_{IMU_2} \\ 1 & : PEF_{IMU_1} < PEF_{IMU_2} \end{cases} \quad (1)$$

Each IMU signal is embedded using the pretrained encoder model and then passed through several Multilayer Perceptrons (MLPs) to predict the binary output. By optimizing the *Binary Cross Entropy Loss*, we can finetune the pretrained feature extractor, enhancing its ability to capture effort information from our exhalation datasets. With these two components combined, the encoder can learn sufficient effort-level information and embed it into vectors, improving the accuracy of the subsequent reconstruction.

Note that in the finetuning stage, the IMU data are undergone similar data augmentation pipeline as discussed in Section 3.3.1, including random shifting in the time domain and noise-adding with $n \sim \mathcal{N}(0,\ 0.1^2)$.

3.3 Maximal Exhalation Reconstruction

Given the audio features from the submaximal exhalation audios and the exhalation effort encoding from the IMU data, in this section, we employ a reconstruction model to predict the ideal maximal-effort breathing patterns. By default, we use five non-maximal exhalations to reconstruct the maximal exhalation pattern to ensure the quality of the reconstruction. This is reasonable since, even in a clinical setup, clinicians always ask patients to perform multiple trials of PFT to ensure the usability of the PFT results [18]. Besides, in our experiments, we observe that most patients are comfortable performing these many casual exhalations. Even so, our system supports a range of 3-5 exhalations as input with zero padding. As discussed in Section 6.2.2, EasySpiro can achieve an acceptable PFT performance even if only three exhalations are provided. The input of the reconstruction model is a 10-channel image tensor composed of five exhalations with two audio channels.

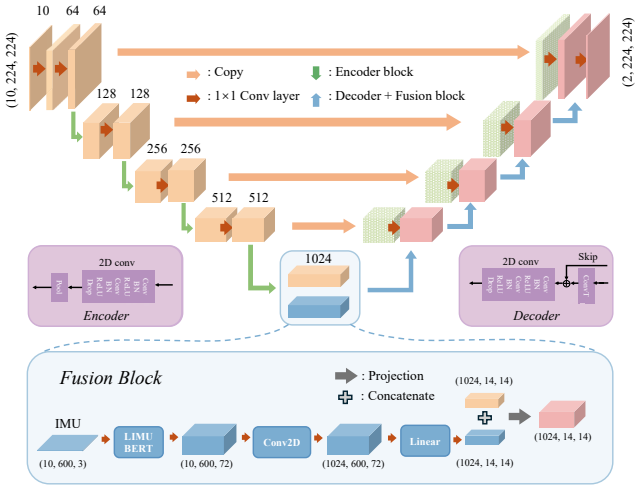


Figure 5: UNet-based reconstruction.

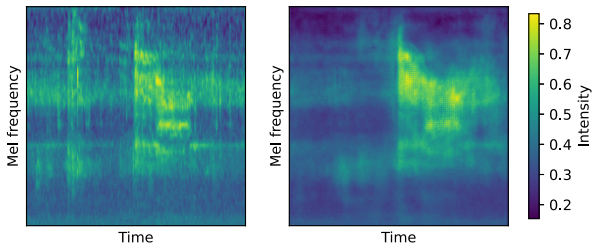


Figure 6: Result of reconstruction. (Left: label; Right: the reconstruction)

3.3.1 Data Augmentation. Before developing the reconstruction model, we use a few data augmentation techniques to enhance its robustness. First, we shuffle the sequence of the ten channels of the input to simulate the scenario where the subject can have any arbitrary sequence when performing the exhalations. Second, we randomly mask two or four channels to simulate the case when the subject only provides three or four exhalations. Third, we randomly shift the spectrogram along the time axis to simulate the cases where the exhalation happens at other time points of the time window. Finally, we randomly add Gaussian noise to the spectrogram following $n \sim \mathcal{N}(0,\ 0.05^2)$.

3.3.2 UNet-based Reconstruction Model. We employ a UNet [45] architecture for our reconstruction model due to its proven effectiveness in image-to-image generation tasks. In the encoding part of the UNet model, the Mel spectrograms and IMU signals are processed separately. The UNet's contracting path only processes the 10-channel spectrograms, utilizing a series of 3×3 convolutional layers followed by max pooling. The previously pretrained feature extractor first encodes the 10-channel IMU signals, projecting the channels and sizes to match the output size of the last encoding layer for the spectrogram. Importantly, we concatenate the IMU information only at the bottom of the UNet, meaning that the IMU features are extracted at a high dimension. An alternative approach would involve adjusting the channel and size of the IMU signals layer by layer and concatenating them with the encoded spectrogram before each max pooling layer, utilizing different levels of IMU data. However, given that IMU data typically contains much less information than audio data, combining them at each contracting layer would dilute the information weight, which is not ideal. Additionally, this approach would significantly increase the computational workload. We will compare these methods in our evaluation in Section 6.

At the bottleneck of the UNet, we get the 1024-channel audio vector and 1024-channel IMU vector with pretrained LIMU-BERT. We concatenate these two vectors to form a 2048-channel vector along with the first dimension. Then we use a series of 3×3 convolutional layers to process this vector to maintain the original 1024-channel size. The expansive path of the UNet is symmetric to the contracting path, with each upsample layer concatenating the corresponding encoding layer's output.

The rationale of the fusion block is that by combining the high-level information from the IMU data with the audio data, we can create a more informative embedding vector with high dimensional representation, which can enhance the efficiency of the subsequent expansive path. Meanwhile, such one-time fusion can reduce the computational complexity on both the encoder and decoder. The entire framework is illustrated in Figure 5.

We can express the process using the formula below as shown in Equation 2. Here, E represents the encoder, F denotes the IMU feature extractor, K is the kernel, f indicates the non-linear mapping, and N is the number of encoding layers. By incorporating the effort representation at the highest dimension, each upsample layer can consider this information and adjust the weights accordingly. The loss function in this reconstruction phase is the mean squared error between the reconstructed spectrogram and the ground truth, in order to minimize the difference between the two.

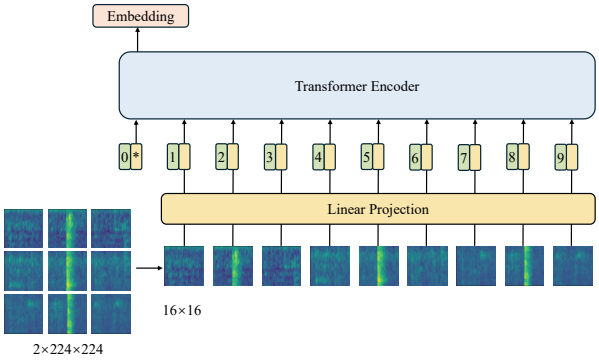


Figure 7: ViT-based prediction.

$$
\begin{aligned}
E_0 &= X \\
E_i &= MaxPool(f(E_{i-1} \cdot K_i + b_i)) \\
E'_N &= Concat(E_N, F(IMU)) \\
X_{up_i} &= Upsample(E_i) \\
Y_{up_i} &= f(X_{up} \cdot K_{up_i} + b_{up_i}) \\
Z_i &= Concat(Y_{up_i}, E_{N-i})
\end{aligned} \tag{2}
$$

The results of the reconstruction are shown in Figure 6. Notably, compared to the original maximal effort breathing on the left, the generated spectrogram exhibits less noise interference. This improvement enhances the system's robustness across various environments.

3.4 PFT Indicators Prediction

In this section, we describe how to make accurate predictions following the reconstruction of the maximum effort breathing spectrogram. We utilize the state-of-the-art Vision Transformer (ViT) [14] as the backbone of our prediction model for its proven record in image processing tasks. Following the design rationale of ViT, we first divide the spectrogram outputted by the previous module into several patches, with each patch treated as a separate vector after linear projection. The structure is illustrated in Figure 7. We then implement a two-phase training strategy that includes a pretraining and fine-tuning phase. During the pretraining phase, the model predicts all eight pulmonary function indicators, assigning equal weight to each during loss calculation and back-propagation. Through this phase, we expect the model to develop a general understanding of pulmonary function prediction, after which we fine-tune it over a few epochs to optimize predictions for each specific indicator by adjusting the corresponding weights. The following section discusses these designs in detail.

3.4.1 Regression Model Architecture. The backbone we adopt is the ViT architecture, a state-of-the-art general vision framework [14]. We choose a transformer-based solution over a

convolutional neural network (CNN) design because it is a better fit for the properties of exhalations.

First, CNNs exhibit a strong inductive bias, particularly in the locality and translation equivariance [64]. This feature enables CNNs to capture local patterns efficiently, which also helps reduce model complexity and training difficulty. However, for the reconstructed maximal effort breathing spectrogram, global information is more critical, as exhalation can last more than 6 to 8 seconds. Thus, CNNs struggle to aggregate over distant audio moments in an image. The second problem is about overfitting. As previously mentioned, our datasets include diverse and inconsistent breathing patterns across subjects, making it possible that the training set may not include samples similar to those in the test set. If our model focuses too much on the locality of the training set, it may only learn features from homogeneous samples, neglecting the sparse ones. Given these considerations, the ViT architecture is more appropriate for the regression tasks involving our heterogeneous datasets.

3.4.2 Pretraining Phase. In the first pretraining phase, we employ a multitask learning strategy based on the transformer backbone. We modify the original ViT model to accommodate our datasets. After feature extraction, we also concatenate personal information with the embedding vector. The 10-channel spectrogram and IMU data combine to form a two-channel image representing maximal effort breathing, from which we need to extract pulmonary function test (PFT) indicators. Although our target indicators are FVC, FEV1, PEF, and FEV1/FVC, medical literature suggests that MEF75, MEF50, MEF25, and MMEF also reflect lung conditions and are included in our datasets. Therefore, we implement an 8-head output as an 8-task learning scheme to predict these eight values simultaneously. During the pretraining phase, we assign equal weight to each output, allowing the model to focus on general lung condition evaluation rather than strengthening the prediction ability for any specific indicator.

Our entire model size is similar to the ViT tiny version; however, we do not utilize any pretrained checkpoints from ImageNet-1k [46] or ImageNet-21k [12]. The data domain of ImageNet, which consists of abundant natural images, is entirely different from that of sound spectrograms. Furthermore, ImageNet primarily focuses on classification tasks, while our goal is to predict accurate absolute values based on the spectrogram. Thus, pretrained checkpoints are not suitable for our scenario.

3.4.3 Finetuning Phase. After the pretraining phase, we proceed to fine-tune the model to predict one specific value at a time for a few epochs. In this step, we set the weight for the current indicator to 1 and the weights for all others to 0.01, allowing the model to concentrate more on the current prediction.

4 Implementation

In this section, we discuss the implementation of EasySpiro. The hardware prototype and the corresponding ground truth spirometer are shown in Figure 9.

Hardware. We separately discuss the implementation of our sensor in two parts: the microphone and IMU. To implement earphones that can be embedded into an earphone, we utilize MAX9813 [13] microphone amplifiers embedded into 3-D printed earphone molds, because they offer very tiny packaging and a low-noise feature. For the audio card, we adopt ALC4032 Serial digital audio solution [11]. The card provides ADC 3.0 / UAC 2.0 protocols and supports 192kHz sample rates, which is sufficient for our scenario. Compared with other papers, our earphone solution can be plugged into mobile or PCs directly without any other converter, making it more convenient. In terms of the IMU sensors, the data acquisition platform consists primarily of an STM32 microcontroller (MCU) and two MPU6050 6-axis IMUs mounted on the back of each of the two headsets. To minimize interference during movement, the control board is fixed to the chest and connected to the IMUs using flexible circuit (FPC) flat cables. Each IMU sample is 50 Hz and timestamps are recorded for synchronization with each sound sample.

Software. As described in Section 3.1, we determine the starting points and extract 12 seconds of data, resulting in one audio channel and three-axis IMU data per ear. Consequently, the input data is structured as two arrays with shapes [(576,000, 2), (600, 6)]. The software of EasySpiro is implemented using Python 3.9. We use Pytorch 1.15 to develop deep-learning models. As for model training, we utilize NVIDIA RTX A6000 GPU, and the optimizer is AdamW [30]. The reconstructed models are trained with 30 epochs. For the indicator prediction model, we first pretrain it using 8 outputs, and then finetune it with the strategy mentioned in Section 3.

Ground truth. Refering to EarSpiro [66], we also use the UBREATH Spirometer System (PF680) [58] as our ground truth collecting device.

5 Dataset Development

In this section, we discuss the dataset development in this research. The development of this dataset is a collaborative effort with a respiratory medical center. This dataset is open-sourced[1].

5.1 Dataset Overview

We recruited 50 participants, including 36 patients with various diseases and 14 subjects with unknown or no disease. Among the patient population, there are 18 COPD or asthma patients, six hypertension patients, and 22 patients with other

[1]Dataset available at https://github.com/ERICXUCHI/EasySpiroDataset

Table 1: Demographics of the subjects.

Stats	COPD/ asthma	Hyper- tension	Before surgery	Others	All
Population	18	6	22	14	50
Age (years)	52.8 (18.8)	62.7 (11.5)	50.8 (13.1)	51.4 (14.3)	51.8 (15.0)
Height (cm)	168.4 (7.0)	164.9 (9.5)	163.7 (10.0)	165.4 (7.5)	165.2 (8.5)
Weight (kg)	64.6 (9.8)	73.8 (17.0)	60.1 (13.1)	66.0 (8.7)	63.6 (12.4)
BMI (kg/m^2)	22.7 (2.7)	26.8 (3.8)	22.3 (3.7)	24.1 (2.9)	23.2 (3.6)
FVC (L)	3.8 (0.7)	3.0 (0.8)	3.3 (0.9)	3.1 (0.8)	3.3 (0.8)
FEV1 (L)	2.6 (0.7)	2.2 (0.4)	2.7 (0.7)	2.5 (0.9)	2.6 (0.7)
PEF (L/s)	6.3 (1.4)	5.6 (1.4)	6.7 (1.6)	6.4 (2.3)	6.4 (1.7)
FEV1/FVC	0.7 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)

Format: mean (standard deviation)

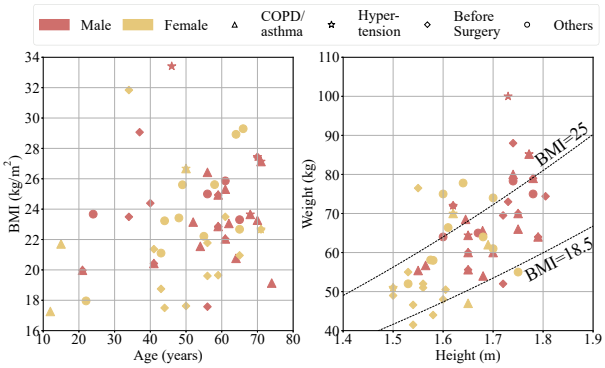
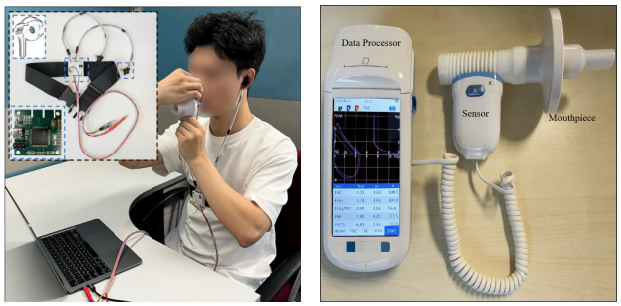


Figure 8: Details of demographics.



(a) EasySpiro hardware (b) Ground truth spirometer

Figure 9: Data collection setup.

diseases who have scheduled surgeries in the subsequent weeks. Note that one patient can have one or more diseases. Each patient contributes around five exhalation maneuvers with various effort levels. Note that the average age of our subjects is larger than 50, at which point one should conduct PFT regularly. We tested our system using this abundant and varied data to validate its feasibility. All COPD patients are diagnosed with mild to moderate severity by senior doctors and examination reports. We also obtained all participants' PFT reports from a hospital medical device to ensure every procedure was valid. Each participant receives 100 CNY[2] compensation, and the whole experiment is approved by the IRB of our institution[3]. All data collection procedures are supervised by the doctors. The demography of the subjects is shown in Table 1 and Figure 8, where the BMI (Body Mass Index) indicates the degree of obesity, FVC (Forced Vital Capacity) reveals the overall lung capacity and FEV1 (Forced Expiratory Volume during the first second) indicates the severity of obstructive lung diseases [21]. The last metric, FEV1/FVC, is an indication of the existence of obstructive lung diseases.

5.2 Data Collection Procedure

The data are collected in a pulmonary function test room in the medical center, which is a semi-open room. Therefore, there would be outside television sounds, queuing machine sounds, and other speech voices. We did not deliberately separate our position from these real scenarios since we cannot guarantee that there is an absolutely silent environment in real use. During data collection, the user needs to wear our customized earphones embedded with microphones and IMU sensors and conduct PFTs. The hardware is shown in Figure 9. First, we recruit the participants and explain the purpose of the study using advertisements and posters, with the help of both doctors and clinical staff. After informed consent and an introduction to this research, we collect the subject's anonymized demographic information, including age, gender, height, weight, and medical history. We also collect their PFT reports from the hospital.

The participants are then asked to wear the earphones and perform PFTs. The participants are instructed to perform the PFTs in a sitting position, with their backs straight and their feet flat on the floor. The participants are also asked verbally to take a deep breath and then exhale as hard and fast as they could into the mouthpiece of the spirometer. During this time, we would not give them suggestions on how to improve the PFT maneuvers, to ensure that they perform the PFTs naturally with different effort levels. At the last attempt, the doctor would instruct them on the standard way of conducting PFT, where they should try their best to inhale and exhale in one breath, and we recorded the maximum effort exhalation after they had learned the standard process. After recording, we would compare each session's PEF index to ensure that the participant did not use maximum effort

[2]100 CNY ≈ 13.7 USD

[3]The Hong Kong University of Science and Technology HREP-2024-0309

Table 2: Comparision with baselines.

Baselines	SpiroSmart [28]		ExhaleSense [42]		EarSpiro [66]		EasySpiro
Dataset*	Original	Ours	Original	Ours	Original	Ours	Ours
FVC	5.2%	22.3%	-†	-	9.9%	19.3%	8.1%
FEV1	4.8%	24.6%	-	-	7.8%	19.8%	7.1%
PEF	6.3%	25.8%	-	-	6.5%	10.3%	4.5%
FEV1/FVC	4.0%	15.4%	7.57%	17.3%	5.1%	16.3%	6.3%

* The original dataset refers to the dataset used in the respective papers. As a comparison, we test the performance of their techniques on our dataset, which contains mostly submaximal exhalation sounds.

† Not provided in the paper.

during the previous sessions. The raw audio sampling rate is 48k Hz, and the raw IMU signal sampling rate is 50 Hz. We record each signal's starting point and align them with the Internet's timestamp, whose error is less than one millisecond.Notably, the noise level in the room measured by a sound level meter is about 60 - 70 dB, which is a typical indoor environment, and the range of breathing sounds spans 42–105 dB, falling within the audio range of earphones.

6 Evaluation

This section will give a detailed evaluation of our system, including a performance study, ablation study, robustness study and demographics study. Note that since our participants' lung conditions vary a lot, we also present our performance at each sub-group in the performance study.

6.1 Evaluation Setup

We first set the evaluation metrics and baselines against which we evaluate the performance of EasySpiro.

6.1.1 Performance Metrics. For all PFTs indicators, we employ *Percentage Error* in Equation 3 as our performance metrics, following the medical standard and previous works [17, 28, 52, 66]. The percentage error is calculated as:

$$Error = \frac{|label - prediction|}{label} \times 100\% \tag{3}$$

6.1.2 Baselines. We select three state-of-the-art studies that predict the PFT indicators based on breathing sounds as our baselines. Since we do not have access to their datasets, we re-implement their methods following the respective papers.

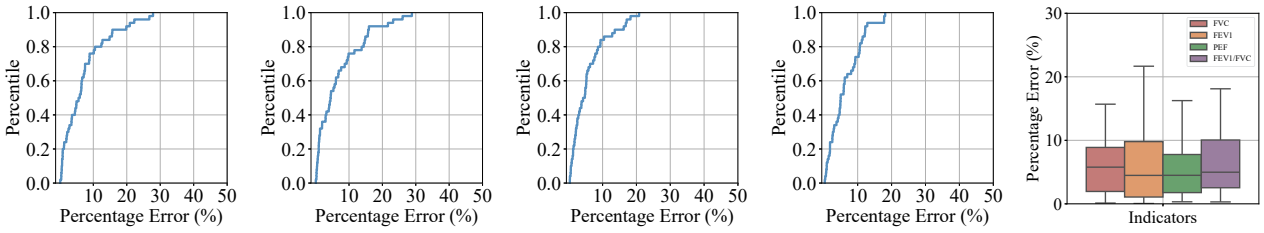
- SpiroSmart [28]. This study outlines key components for feature extraction, such as envelope detection, spectrogram processing, and linear predictive coding (LPC). After processing these features, machine learning regression predicts all four lung indicators.
- ExhaleSense [42]. This research focuses on detecting exhalation and extracting waveform features to predict lung obstruction parameters, specifically FEV1/FVC. After identifying the exhalation phase, the study computes both temporal and spectral features from the envelope signals, followed by the machine learning regularized regression model.
- EarSpiro [66]. The authors first apply STFT and Mel filter bank to the raw audio signals. They then utilize the energy profile of the Mel spectrogram to segment the expiration phase. After selecting the corresponding period, they design a CNN-GRU-based model for PFT measurement.

Note that while their original datasets consist of maximal effort breathing sounds, our dataset contains mostly submaximal effort exhalation sounds. The performance of these baselines on their original datasets and on our datasets are shown in Table 2.

6.2 Performance Study

In this section, we evaluate EasySpiro's general functionality. We first evaluate EasySpiro's performance in predicting the four target lung function indicators. Then, we test EasySpiro's performance when different numbers of exhalation samples are used.

6.2.1 Overall Performance. We use leave-one-subject-out (LOSO) validation in this evaluation. The overall performance is shown in Figure 10. The average percentage error for FVC, FEV1, PEF, and FEV1/FVC is 8.08%, 7.12%, 4.50%, and 6.35%. The indicator errors are a little bit higher than similar works [52, 66]. This is because they require the subjects to perform maximal effort breathing while we just adopt submaximal effort in our scenario, a much easier and more comfortable way to predict lung conditions. Furthermore, we validate our system on various patients, including those with COPD, asthma, high blood pressure, and other diseases. A more detailed breakdown analysis will be shown later in this section. Notably, medical-grade devices are generally designed to tolerate a percentage error of up to 5% as documented in various studies and standards [3, 10, 18]. However, an around 7% error rate has also been observed, which is



(a) FVC (b) FEV1 (c) PEF (d) FEV1/FVC (e) Overall

Figure 10: Overall performance. (a)-(d) CDF plots. (e) Box plot.

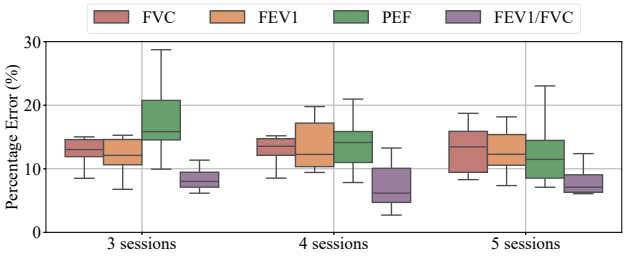


Figure 11: Impact of exhalation sessions.

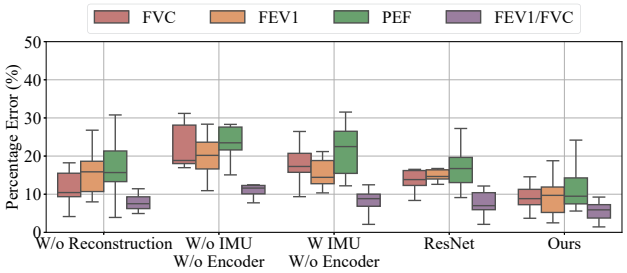


Figure 12: Ablation study.

consistent with the findings of previous research, as demonstrated in Table 2.

6.2.2 Performance with Different Session Numbers. As we mention in Section 3.3.1, even though most patients are comfortable with conducting five PFT maneuvers, our system needs to be robust for fewer sessions. To validate its performance, we evaluate our system on three, four, and five sessions separately. This result employs 10-fold cross-validation. The comparison of each case is present in Figure 11. The results show that even if we only utilize three sessions of exhalation for each person, the percentage error is still acceptable. The reason is that during our reconstruction phase, we augment enough samples with three, four, and five valid sessions by masking other channels. In this way, our reconstruction model is empowered with a robust ability to generate the maximal effort exhalation spectrogram even with fewer breathing numbers.

6.2.3 System Latency. The following details outline the system latency, as presented in Table 3. On the CPU, the total processing time is approximately 3 seconds, whereas the GPU achieves a significantly reduced total time of about 0.5 seconds. However, since PFT is not a real-time task, moderate delays are tolerable and acceptable, rendering the overall system latency satisfactory.

Table 3: System delay (in ms)

Device	Proc		Recon	Pred	Total
CPU	1101.78		990.51	625.30	2717.59
	AUD	**IMU**			
	474.44	627.34			
GPU	471.00		4.74	8.54	484.28
	AUD	**IMU**			
	465.07	5.93			

Proc: Audio and IMU pre-processing. Recon: maximal effort reconstruction. Pred: PFT indicator prediction.

6.3 Ablation Study

In this section, we present a comprehensive ablation study to verify the contribution of each component in our proposed model. The technique modules are divided into three parts. First, we validate the function of spectrogram reconstruction by using only raw, submaximal data to predict the pulmonary function indicators. Second, since we claim that we observe the IMU would contribute to containing effort information and leverage a self-supervised encoder to extract features from the signals, we remove the specially-designed IMU encoder or remove the IMU modality completely. Third, since we use ViT as our backbone instead of a conventional CNN-based network, we also compare these two frameworks' performance. All results in this section adopt 10-fold cross-validation.

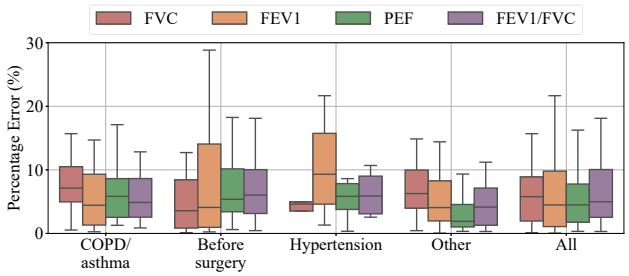


Figure 13: Demographic study - disease.

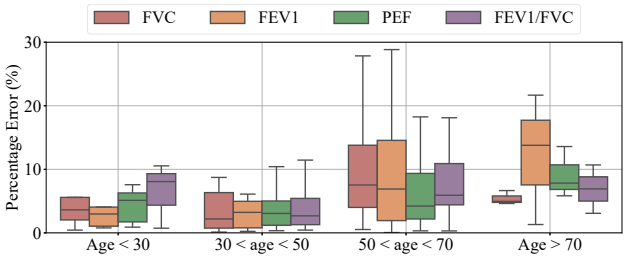


Figure 14: Demographic study - age.

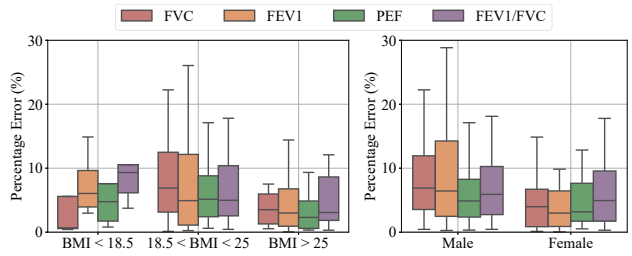


Figure 15: Demographic study - BMI and gender.

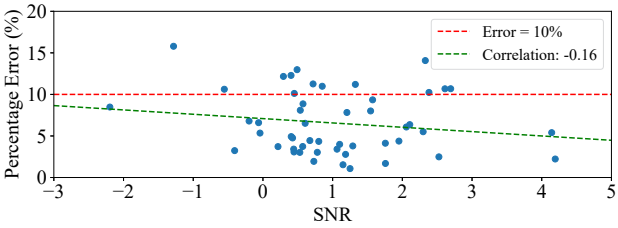


Figure 16: Impact of noise.

6.3.1 Impact of Spectrogram Reconstruction. Since we propose a two-phase learning scheme by first reconstructing the maximal effort exhalation spectrogram and then predicting the lung condition based on the previous result, we need to verify the effectiveness of the reconstruction module. The comparison is shown in Figure 12. We can observe that without reconstruction, the performance drops to a certain degree. This is because, through the intermediate reconstruction process, we add a "supervised loss" by computing the difference between the generative maximal effort breathing spectrogram and real ones. By this means, we can provide more information to guide the model in generating a proper spectrogram and ultimately contribute to the prediction of the final indicators. Additionally, this method is aligned with the medical process, which increases the interpretability of our design.

6.3.2 Impact of IMU. As we mentioned previously, we observe that the effort level from IMU signals would help the final prediction. Thus, we systematically remove the IMU modality or keep the signal but remove the encoder to validate its contribution. The results are presented in the second and third columns of Figure 12. It shows that without the IMU modality at all, the PFT indicator prediction will have large errors with the average error rate exceeding 20%. With the introduction of IMU, the error rate is reduced. Further, the error rate is reduced significantly when using our IMU encoder trained by the self-supervised pipeline. This result proves the effectiveness of introducing IMU to guide the PFT prediction and the superiority of our specially designed IMU encoding method. A separate evaluation in terms of the spectrogram reconstruction performance is presented in Section 6.6.

6.3.3 Impact of ViT Architecture. We choose ViT because of its great ability for global feature fusion. However, previous works also claim for limited data, ResNet [20] may perform better. Therefore, in this part, we compare the ResNet-based architecture versus the ViT-based architecture. The comparison is also shown in Figure 12. This helps to prove that our model could extract global information from the spectrogram more effectively. Moreover, since it's easy for ResNet to overfit, our model also shows the potential for various data distribution.

6.4 Demographic Study

The demographics of our datasets are presented in Section 5. This section demonstrates our system performance on different demographic groups. We evaluate our system in terms of health condition, gender, age, and BMI value. Specifically, the subjects are grouped into underweight (BMI<18.5), normal (18.5<BMI<25), and overweight (BMI>25) according to the international standard.

6.4.1 Subjects with Diseases. We first analyze the performance of EasySpiro on the patient population in our dataset, including COPD, asthma, hypertension patients, and other patients with surgery scheduled in the subsequent weeks. The result is shown in Figure 13. Since the COPD and asthma

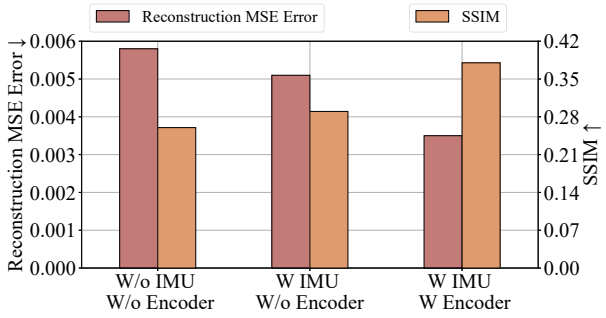


Figure 17: Spectrogram reconstruction performance.

patients' lungs are obstructive more or less, their breathing sounds are weaker than others. Therefore, the error in this user group is slightly larger than that of others.

6.4.2 Age. The impact of different ages on the indicators prediction is shown in Figure 14. It suggests that the performance of the elderly is worse than that of the young. This is because most patients are concentrated in the aged subjects, and the aged people often have weaker breathing sounds, which are hard to detect. Moreover, since the elderly always have insufficient muscle strength, their IMU signals are not so distinguishable compared with others, and their own tremors could also affect the IMU sensors.

6.4.3 Gender and BMI. We also evaluate the performance in terms of gender variation. The results are shown in Figure 15. We observe that there is a slight decline in performance among the male subjects. Since most lung obstructive patients are male, their overall predictions are not as accurate as those of the female subjects.

For the BMI values, we separated all the participants into three categories: underweight, normal, and overweight. The details are displayed in Figure 15. We find that compared with that of normal and underweight subjects, our system tends to achieve higher accuracy on overweight subjects. We analyze that most lung disease patients are very thin, thus their BMI values should be below normal degree.

6.5 Robustness Study

The most important environmental factor in our scenario is the ambient noise. As we mention in Section 3.3.2, the generative maximal breathing spectrogram tends to exhibit less noise interference. To verify our system's robustness, we divide our datasets into four environments, which are a quiet room, a room with an air conditioner, a room with a television on, and a room with others speaking. The relationship between SNR (Equation 4) and percentage error is shown in Figure 16.

$$SNR = 10 \cdot log_{10}(P_s/P_n) \tag{4}$$

where P_s is the signal power and P_n is the noise power.

From this analysis, we could conclude that our system can be resilient to environmental noise in the real world, and the performance is almost unaffected.

6.6 Spectrogram Reconstruction Performance

The metrics to measure its function are reconstruction loss and the Structural Similarity Index (SSIM) [62], where we compare the generative maximal spectrogram with the original spectrogram. SSIM is a metric used to assess the quality of digital images by measuring the similarity between two images, and the formula is written as follows

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c2)} \tag{5}$$

where μ is the mean of the image, σ is the standard deviation of the image, and c is a constant. If $SSIM$ is higher, the two images are more similar to each other. The evaluation result is shown in Figure 17. This result indicates that the introduction of IMU enhances the reconstruction performance, and the use of the IMU encoder trained by self-supervised learning further boosts the performance. This evaluation result is aligned with that of Section 6.3.1.

7 Related Work

In this section, we review the research related to this paper. They fall into two categories: mobile spirometry systems and earphone-based sensing systems.

7.1 Mobile Spirometry Systems

Recent studies indicate that spirometry can be conducted using mobile devices, yielding reliable results. SpiroSmart [28] was the first to utilize a mobile phone microphone to measure lung function. Several other studies follow a similar approach, using the sound of breathing combined with maximum effort to conduct pulmonary function tests [17, 42, 56, 57, 66].

Some research focuses on analyzing the motion of breathing under maximum effort. SpiroSonic [52] employs acoustic sensors to convert chest wall motion into lung function indices. SpiroFi [72] captures chest wall movement from variations in WiFi signals and interprets this data into lung function indices. Han *et al.* [19] use millimeter-wave radar for contactless sensing of chest and abdominal motion, recovering Expiratory Volume (EV) curves from the overall motion data. Similarly, mmFlow [1] also employs millimeter-wave technology to analyze the subtle vibrations produced by airflow when individuals breathe on the device's surface.

Several studies explore the feasibility of creating a 3-D model to predict lung conditions. Kaiser *et al.* [25] designed a set of vortex whistles that generate sound frequencies

proportional to airflow speed, allowing the estimation of the flow-volume (F-V) curve by analyzing the frequency of recorded expiratory sounds. Yin *et al.* [70, 71] developed a mouthpiece with a specific airway tube, reconstructing the human airway profile through analysis of reflected acoustic waves captured by a smartphone microphone, and extracting features to assess lung function.

Additionally, some research utilizes cough, speech, or other biomarkers to estimate lung health indicators [37–39, 47, 48, 60, 67]. MMLung [36] integrates audio signal data from multiple modalities and tasks to achieve notable performance in lung function estimation. Cheng *et al.* [9] utilize a cell phone's motion sensor to classify GOLD (Global Initiative for Chronic Obstructive Lung Disease) levels. However, these solutions do not adhere to the gold standard PFT and are unable to measure peak expiratory flow (PEF) indicators. In addition, there is a parallel study that tries to perform PFT via natural speech sounds [8].

7.2 Earphone-based Sensing Systems

Headphones are lightweight devices, and many earphone-based systems have been developed. These systems can be categorized into three main areas: (i) health monitoring, (ii) human-computer interaction (HCI), and (iii) authentication and identification.

Several earphone-enabled health monitoring systems have emerged. Martin *et al.* [33] and Butkow *et al.* [7] utilize in-ear microphones to measure heart rates by analyzing audio features. Wang *et al.* [61] employed earbuds to detect breathing phases, aiming to reduce user distraction. The eBP system [6] measures blood pressure using PPG sensors. Ferlini [16] demonstrated that in-ear PPG can accurately detect vital signs, including heart rate (HR), heart rate variability (HRV), blood oxygen saturation, and respiration rate (RR). For disease-specific monitoring, EarHealth [24] evaluates hearing health by analyzing ear canal geometry and eardrum mobility. Additionally, EarWalk [23] uses common wireless wearables to provide continuous feedback on gait changes, helping to reduce knee stress.

In the field of HCI, some studies focus on recognizing facial expressions [29, 51]. OESense is an acoustic-based in-ear system for motion sensing, including step counting and activity recognition. TeethTap [54] and EarSense [41] use earphones to detect teeth movement for human-machine interaction. JawSense [26] decodes unvoiced phonemes for hands-free, privacy-preserving interaction.

For authentication and identification, Xie *et al.* [68] introduced TeethPass, which uses earbuds to collect occlusal sounds from binaural canals for authentication. ToothSonic [63] extracts multi-level acoustic features that reflect intrinsic toothprint information, aiding in the authentication.

8 Discussion

In this section, we will discuss the limitations and potential extensions of EasySpiro.

Large-scale deployment on COPD patients. While our system generally exhibits strong performance, its accuracy decreases slightly when utilized with COPD patients. Since the lungs of COPD patients are partially obstructed, their maximal exhalation flow is slower and the sound is weaker compared to that of healthy individuals. As a result, some of their non-maximal breathing sounds may be inaudible. This leads to a decline in accuracy when our system is used exclusively for COPD patients. Therefore, future works shall recruit a larger COPD population and conduct a large-scale deployment to further evaluate the in-the-wild performance of the design.

Inspiratory measurement. While the exhalation phase provides more information than the inhalation phase, there are also helpful indices such as Peak Inspiratory Flow (PIF) and Forced Inspiratory Flow at 50% of Vital Capacity (FIF50). However, our system focuses only on the exhalation phase because the sound of inhalation is extremely weak and often indistinguishable, especially for COPD patients. Even when our microphones are placed in the ear, the waveform can be buried in ambient noise in real-world settings. Therefore, we omit inspiratory measurements and provide lung condition indicators based solely on the expiratory phase.

9 Conclusion

This paper proposes EasySpiro, the first mobile spirometer system built from commercial-off-the-shelf microphones and IMU sensors to predict lung condition indicators based on arbitrary, non-maximal breathing effort. Particularly, EasySpiro utilizes UNet-based spectrogram reconstruction techniques to generate maximal breathing from submaximal ones enhanced by self-supervised IMU encoding. Then the system leverages the state-of-the-art vision transformer backbone to predict the lung function indicators. We collaborate with a medical center and build a dataset that includes 50 subjects with various health conditions to validate our system. The experimental results show that EasySpiro achieves high accuracy in real-world scenarios. Additionally, we open-source the first PFT dataset with all real patients' exhalation sounds.

Acknowledgments

We are very grateful to the reviewers and the shepherd for their insightful comments. We also thank Jinjian Wang and Guanting Lin for their help in developing the hardware prototype for this research. This research is supported in part by the Hong Kong Research Grants Council (RGC) under Contracts CERG 16206122, 16204523, AoE/E-601/22-R, R6021-20, and Contract R8015.

References

[1] Aakriti Adhikari, Austin Hetherington, and Sanjib Sur. 2021. mmFlow: Facilitating At-Home Spirometry with 5G Smart Devices. In *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[2] Tousif Ahmed, Md Mahbubur Rahman, Ebrahim Nemati, Mohsin Yusuf Ahmed, Jilong Kuang, and Alex Jun Gao. 2023. Remote breathing rate tracking in stationary position using the motion and acoustic sensors of earables. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.

[3] Richard B Belzer and R Jeffrey Lewis. 2019. The practical significance of measurement error in pulmonary function testing conducted in research settings. *Risk Analysis* 39, 10 (2019), 2316–2328.

[4] Dolores Blanco-Almazán, Willemijn Groenendaal, Lien Lijnen, Rana Önder, Christophe Smeets, David Ruttens, Francky Catthoor, and Raimon Jané. 2022. Breathing pattern estimation using wearable bioimpedance for assessing COPD severity. *IEEE Journal of Biomedical and Health Informatics* 26, 12 (2022), 5983–5991.

[5] 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.

[6] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[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 *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 200–209.

[8] Yetong Cao, Dong Ma, Wentao Xie, Qian Zhang, and Jun Luo. 2025. ESPIRO: Natural Pulmonary Function Monitoring via Earphone-Acquired Speech. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking* (Hong Kong, China) *(MobiCom '25)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[9] Qian Cheng, Joshua Juen, Shashi Bellam, Nicholas Fulara, Deanna Close, Jonathan C Silverstein, and Bruce Schatz. 2017. Predicting pulmonary function from phone sensors. *Telemedicine and e-Health* 23, 11 (2017), 913–919.

[10] B Cooper and A Butterfield. 2009. Quality control in lung function testing. *ERS Buyers' Guide Respir Care Prod* (2009), 24–38.

[11] THITRONIX TECHNOLOGY CORP. 2024. ALC4032 Serial digital audio solution. https://www.thitronix.com/En/Products/List-4-14-0-1.html Accessed Aug 26, 2024.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[13] Analog Devices. 2024. MAX9813 Microphone. https://www.analog.com/en/products/max9813.html#part-details Accessed Aug 26, 2024.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[15] AI Dyachenko, GA Lyubimov, IM Skobeleva, and MM Strongin. 2011. Generalization of the mathematical model of lungs for describing the intensity of the tracheal sounds during forced expiration. *Fluid Dynamics* 46, 1 (2011), 16–23.

[16] Andrea Ferlini, Alessandro Montanari, Chulhong Min, Hongwei Li, Ugo Sassi, and Fahim Kawsar. 2021. In-ear PPG for vital signs. *IEEE Pervasive Computing* 21, 1 (2021), 65–74.

[17] Mayank Goel, Elliot Saba, Maia Stiber, Eric Whitmire, Josh Fromm, Eric C Larson, Gaetano Borriello, and Shwetak N Patel. 2016. Spirocall: Measuring lung function over a phone call. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5675–5685.

[18] 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.

[19] Shijie Han, Dongheng Zhang, Jinbo Chen, Haoyu Wang, Jinli Zhang, Qibin Sun, and Yan Chen. 2023. Fine-grained Lung Function Sensing based on Millimeter-Wave Radar. In *2023 International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 471–476.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] The Johns Hopkins Hospital. 2024. Pulmonary Function Tests. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/pulmonary-function-tests Accessed Aug 26, 2024.

[22] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[23] Nan Jiang, Terence Sim, and Jun Han. 2022. EarWalk: towards walking posture identification using earables. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications*. 35–40.

[24] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. Earhealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th annual international conference on mobile systems, applications and services*. 397–408.

[25] Spencer Kaiser, Ashley Parks, Patrick Leopard, Charlie Albright, Jake Carlson, Mayank Goel, Damoun Nassehi, and Eric C Larson. 2016. Design and learnability of vortex whistles for managing chronic lung function via smartphones. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 569–580.

[26] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. JawSense: recognizing unvoiced sound using a low-cost ear-worn system. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 44–49.

[27] Vladimir I Korenbaum, Irina A Pochekutova, Anatoly E Kostiv, Veronika V Malaeva, Maria A Safronova, Oksana I Kabantsova, and Svetlana N Shin. 2020. Human forced expiratory noise. Origin, apparatus and possible diagnostic applications. *The Journal of the Acoustical Society of America* 148, 6 (2020), 3385–3391.

[28] 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 *Proceedings of the 2012 ACM Conference on ubiquitous computing*. 280–289.

[29] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[30] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[31] Shuyi Ma, Ariel Hecht, Janos Varga, Mehdi Rambod, Sarah Morford, Shinichi Goto, Richard Casaburi, and Janos Porszasz. 2010. Breath-by-breath quantification of progressive airflow limitation during exercise

in COPD: a new method. *Respiratory medicine* 104, 3 (2010), 389–396.

[32] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*. 49–58.

[33] Alexis Martin and Jérémie Voix. 2017. In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring. *IEEE Transactions on Biomedical Engineering* 65, 6 (2017), 1256–1263.

[34] MJ Morris and DJ Lane. 1981. Tidal expiratory flow patterns in airflow obstruction. *Thorax* 36, 2 (1981), 135–142.

[35] MJ Morris, RG Madgwick, I Collyer, F Denby, and DJ Lane. 1998. Analysis of expiratory tidal flow patterns as a diagnostic tool in airflow obstruction. *European respiratory journal* 12, 5 (1998), 1113–1117.

[36] Mohammed Mosuily, Lindsay Welch, and Jagmohan Chauhan. 2023. MMLung: moving closer to practical lung health estimation using smartphones. (2023).

[37] Viswam Nathan, Korosh Vatanparvar, Md Mahbubur Rahman, Ebrahim Nemati, and Jilong Kuang. 2019. Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices. In *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–4.

[38] 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 *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 4491–4497.

[39] Ebrahim Nemati, Xuhai Xu, Viswam Nathan, Korosh Vatanparvar, Tousif Ahmed, Md Mahbubur Rahman, Dan McCaffrey, Jilong Kuang, and Alex Gao. 2022. Ubilung: Multi-modal passive-based lung health assessment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 551–555.

[40] Masafumi Nozoe, Kyoshi Mase, Shigefumi Murakami, Makoto Okada, Tomoyuki Ogino, Kazuhiro Matsushita, Sachie Takashima, Noriyasu Yamamoto, Yoshihiro Fukuda, and Kazuhisa Domen. 2013. Relationship between spontaneous expiratory flow-volume curve pattern and airflow obstruction in elderly COPD patients. *Respiratory Care* 58, 10 (2013), 1643–1648.

[41] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[42] Md Mahbubur Rahman, Tousif Ahmed, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, Erin Blackstock, and Jilong Kuang. 2020. Exhalesense: Detecting high fidelity forced exhalations to estimate lung obstruction on smartphones. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[43] Harpreet Ranu, Michael Wilde, and Brendan Madden. 2011. Pulmonary function tests. *The Ulster medical journal* 80, 2 (2011), 84.

[44] Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[47] 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 *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[48] Keum San Chun, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Md Mahbubur Rahman, Erin Blackstock, and Jilong Kuang. 2020. Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[49] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[50] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors* 14, 6 (2014), 10146–10176.

[51] Xingzhe Song, Kai Huang, and Wei Gao. 2022. Facelistener: Recognizing human facial expressions via acoustic sensing on commodity headphones. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 145–157.

[52] 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 *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. Association for Computing Machinery, New York, NY, USA, Article 52, 14 pages.

[53] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.

[54] Wei Sun, Franklin Mingzhe Li, Benjamin Steeper, Songlin Xu, Feng Tian, and Cheng Zhang. 2021. Teethtap: Recognizing discrete teeth gestures using motion and acoustic sensing on an earpiece. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 161–169.

[55] Karl Peter Sylvester, Nigel Clayton, Ian Cliff, Michael Hepple, Adrian Kendrick, Jane Kirkby, Martin Miller, Alan Moore, Gerrard Francis Rafferty, Liam O'Reilly, et al. 2020. ARTP statement on pulmonary function testing 2020. *BMJ Open Respiratory Research* 7, 1 (2020), e000575.

[56] Tharoeun Thap, Heewon Chung, Changwon Jeong, Ki-Eun Hwang, Hak-Ryul Kim, Kwon-Ha Yoon, and Jinseok Lee. 2016. High-resolution time-frequency spectrum-based lung function test from a smartphone microphone. *Sensors* 16, 8 (2016), 1305.

[57] Hai Anh Tran, Quynh Thu Ngo, and Huy Hoang Pham. 2015. An application for diagnosing lung diseases on Android phone. In *Proceedings of the 6th International Symposium on Information and Communication Technology*. 328–334.

[58] UBREATH. 2024. UBREATH Spirometer System (PF680). https://www.e-linkcare.com/spirometer-system-pf680-product/ Accessed Aug 26, 2024.

[59] 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.

[60] 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 *2021 43rd*

Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 7237–7243.

[61] Zihan Wang, Tousif Ahmed, Md Mahbubur Rahman, Mohsin Y Ahmed, Ebrahim Nemati, Jilong Kuang, and Alex Gao. 2022. Real-Time Breathing Phase Detection Using Earbuds Microphone. In 2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN). 1–4.

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[63] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[64] Zihao Wang and Lei Wu. 2024. Theoretical analysis of the inductive biases in deep convolutional networks. *Advances in Neural Information Processing Systems* 36 (2024).

[65] EM Williams, Tom Powell, M Eriksen, P Neill, and R Colasanti. 2014. A pilot study quantifying the shape of tidal breathing waveforms using centroids in health and COPD. *Journal of clinical monitoring and computing* 28 (2014), 67–74.

[66] Wentao Xie, Qingyong Hu, Jin Zhang, and Qian Zhang. 2023. EarSpiro: Earphone-based Spirometry for Lung Function Assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–27.

[67] Wentao Xie, Chi Xu, Yanbin Gong, Yu Wang, Yuxin Liu, Jin Zhang, Qian Zhang, Zeguang Zheng, and Shifang Yang. 2024. DeepBreath: Breathing Exercise Assessment with a Depth Camera. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 137 (Sept. 2024), 26 pages.

[68] Yadong Xie, Fan Li, Yue Wu, Huijie Chen, Zhiyuan Zhao, and Yu Wang. 2022. TeethPass: Dental occlusion-based user authentication via in-ear acoustic sensing. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1789–1798.

[69] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems* (Coimbra, Portugal) *(SenSys '21)*. Association for Computing Machinery, New York, NY, USA, 220–233.

[70] Xiangyu Yin, Kai Huang, Erick Forno, Wei Chen, Heng Huang, and Wei Gao. 2022. Out-Clinic Pulmonary Disease Evaluation via Acoustic Sensing and Multi-Task Learning on Commodity Smartphones. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 1182–1188.

[71] 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 *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 110–123.

[72] 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 *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*. IEEE, 1–10.

[73] Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. 2024. Towards Open Respiratory Acoustic Foundation Models: Pretraining and Benchmarking. *arXiv preprint arXiv:2406.16148* (2024).

[74] Quan Zhou, Jianhua Shan, Wenlong Ding, Chengyin Wang, Shi Yuan, Fuchun Sun, Haiyuan Li, and Bin Fang. 2021. Cough recognition based on mel-spectrogram and convolutional neural network. *Frontiers in Robotics and AI* 8 (2021), 580080.