ExhaleSense: Detecting High Fidelity Forced Exhalations to Estimate Lung Obstruction on Smartphones

Md Mahbubur Rahman*, Tousif Ahmed*, Ebrahim Nemati*, Viswam Nathan*, Korosh Vatanparvar*, Erin Blackstock†, Jilong Kuang*
*Digital Health Lab, Samsung Research America, Mountain View, CA, USA
†Pulmonary and Critical Care Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
m.rahman2@samsung.com

***Abstract*—Spirometry is the gold standard to measure lung functions by estimating the maximum air an individual can forcefully exhale as quickly as possible. It is used not only to diagnose lung diseases such as asthma, chronic obstructive pulmonary disease (COPD) but also to assess the severity of the pulmonary condition. However, spirometry requires a specialized device called spirometer, which is mostly available in clinical facilities and cumbersome to use. Recent works have shown the feasibility of using smartphone microphone to estimate lung functions from forced exhalation effort sounds. However, maintaining the fidelity of lung function estimation on smartphones becomes challenging in unsupervised field environment in presence of other sounds such as coughs, deep inhalation, regular breathing, and speech. In this paper, we present ExhaleSense that detects forced exhalation efforts on smartphones from audio time-series data, distinguishes high fidelity efforts from poor efforts, and estimates lung obstruction. By conducting three studies with 211 pulmonary patients and healthy subjects, we show that ExhaleSense can detect forced exhalation sounds with 96.74% F1-score and estimate lung obstruction with mean absolute error as low as 7.57%. ExhaleSense shifts the gear of smartphone spirometry research from feasibility to ensuring effort quality towards high fidelity lung function estimation in unsupervised field settings.**

***Index Terms*—mobile spirometry, pulmonary patient, audio, data quality**

I. Introduction

Despite significant advances in health care, chronic respiratory diseases (CRDs) are the third leading causes of death in the world for the last 20 years [1]. Chronic respiratory diseases refer to a group of diseases, primarily chronic obstructive pulmonary disease (COPD) and asthma, affect the airways, and cause difficulty in breathing. Spirometry is the most common way to diagnose the respiratory diseases, which measures the lung condition by estimating the speed and amount of airflow of an individual. Spirometry tests are conducted using a Spirometer in clinical facilities under the supervision of a skilled technician and can be challenging in low and middle-income countries[1] as the test requires expensive equipment, human and financial resources, and technical support [3].

To make the spirometry more accessible and cost-effective, several works such as MobiSpiro [12], SpiroSmart [13], SpiroCall [8], SpiroConfidence [10] showed the feasibility of mimicking the spirometry test on smartphones. For example, Larson et al., [13] developed a smartphone application called SpiroSmart and showed the possibility to measure lung functions using audio data captured by smartphone microphones. Previous works show the promise and potential of using smartphone as a consumer-grade alternative to measure lung condition and make the assessment available anywhere, anytime. However, when the users perform the spirometry task on a smartphone at home without clinician supervision, they oftentimes do not push their effort to achieve their maximum lung capacity that can produce nonsensical results that are not representative of their lung health [9]. Moreover, variations of the interactions between phone and the user at home also affect the exhalation sounds and assessment quality. For example, the distance between the device and the mouth and the device orientation affect the audio quality, which could yield inaccurate results. Therefore, it is critical to consider the variation in human device interactions associated with unsupervised, in-home assessment scenarios to detect the high fidelity forced exhalation efforts on smartphones.

A few prior researches took steps to assess the quality of spirometry efforts. Melia et al. [11], detected errors in spirometry efforts based on the flow-volume curves captured from a traditional spirometer. Viswanath et al. [10], classified valid spirometry efforts captured using smartphone in a clinical setting, given that the forced exhalation sounds are already segmented. However, while monitoring at home, the forced exhalation sound segments need to be pin-pointed in continuous audio time series data and then, the fidelity of the efforts need to be assessed by controlling the quality parameters associated with unsupervised scenarios mentioned above. In this paper, we present *ExhaleSense*, a smartphone based approach to automatically detect forced exhalation sounds in a continuous audio stream captured by its built-in microphones, assess the quality of the effort, and estimate the lung condition.

To train and evaluate the performance of ExhaleSense, we have conducted three independent studies in collaboration with Brigham and Women's Hospital (BWH), one of the largest teaching hospitals of Harvard Medical School, totaling 211 participants including chronic pulmonary patients and healthy

[1]More than 90% COPD related death occurs in low and middle-income countries [2].

978-1-7281-4657-7/20/$31.00 ©2020 IEEE

controls. In the first study, we have collected data from 131 subjects (91 patients and 40 healthy) in a usability lab at Samsung Research America (SRA) where we have recorded forced exhalation sound on Samsung Note8 smartphones. The ground truth lung function measurements were captured by an FDA approved portable spirometer, called GoSpiro [15]. We have conducted the second study in BWH with 70 participants (60 patients and 10 healthy) where each subject was supervised by expert clinicians to perform the test on a hospital-grade spirometer. The participants were also requested to blow their exhalation into the phone. In the third study conducted in BWH, we have collected data from 10 participants who were given the medicine as a test to collect controlled deterioration of their pulmonary condition and the smartphone spirometry after each dose of medication. In all three studies, we also collected lung health estimates from spirometer as a gold standard ground truth. An example of our tests is depicted in the left side of the Figure 1.

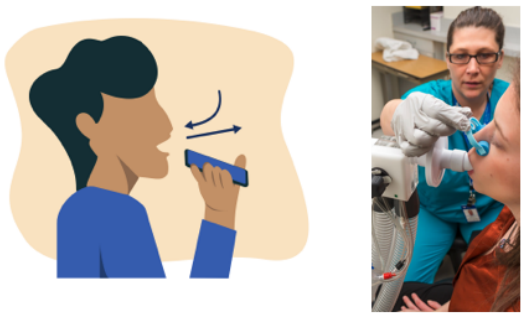


Fig. 1: Subject holding the phone and performing the forced exhalation as fast as possible on a smartphone microphone (left) and using a clinical spirometer (right)

Experiments on the above dataset show that our model can detect can detect forced exhalation with 96.74% F1-score and estimate the lung obstruction bio-marker called FEV1/FVC ratio with 7.57% mean absolute error with respect to the gold-standard spirometers. We further collected usability data from the participants to compare their spirometry experience with a smartphone and a spirometer. Our experiment reveals that the patients show higher preference on using smartphone spirometry over the portable spirometer and the clinical spirometer. Our approach makes the next stride to bridge the gap between the feasibility of the smartphone spirometry and the reality of its deployment in patient community.

In short, our contributions are the following:

- Automatically detect forced exhalation sound on a smartphone with 96.74% F1-score to pinpoint a forced exhalation segment in an audio time series data.
- Assess the quality of the effort and present explainable quality metrics for smartphone spirometry informed by American Thoracic Society guidelines.
- Present three independent studies with chronic pulmonary patients and show methods to rigorously annotate pulmonary audio dataset to develop reliable models.
- Estimate lung obstructions by considering both initial one second features and whole forced exhalation segment features to with 7.57% mean absolute error (MAE).
- Compare the usability of the phone spirometry with respect to a hospital grade spirometer and a connected portable spirometer.

II. Background and Related Works

A. Traditional Lung Function Measurements

Spirometry is the most widely used standard test to measure lung function. A *Spirometer* is a device with a mouthpiece hooked up with a big machine. Patients need to insert the spirometer tube into the mouth and attach a nose clip during the test. It measures flow rate of air as it passes through a mouthpiece. During the test, participants takes a deep breath and then exhales the air as fast as possible and as long as possible (right picture of the Figure 1). The spirometer measures the speed and volume of the airflow and computes several lung function metrics. Three of the most widely used lung function measures are the following:

- **Forced Vital Capacity (FVC):** This is the amount of air an individual can exhale quickly forcefully after a deep inhalation.
- **Forced Expiratory Volume in one second (FEV1):** This is the amount of air expired during the initial one second of the forced exhalation.
- **FEV1/FVC ratio:** This is the ratio of FEV1 and FVC values. This is the most common lung bio-marker that indicates the presence of airflow obstruction in chronic lung patients such as COPD and asthma. In this paper, our model predicts this number based on forced exhale data captured on a smartphone microphone.

B. Pulmonary Assessment Using Mobile Devices

Mobile sensor-based pulmonary assessment methods [4], [5], [8], [13] are more related to our work. Juen et al. [5] and Cheng et al., [6] demonstrated that monitoring of natural walk during daily activities using the smartphone inertial sensors could be useful for predicting lung function in cardiopulmonary patients. Infante et al., [7] investigated the use of cough sound recorded by a smartphone for the screening and diagnosis of pulmonary disease.

The closest related works are those that aim to develop smartphone-based spitometry sensing methods. Larson et al. [13] developed a system called SpiroSmart that can measure lung function using a smartphone device. SpiroSmart requires a user to hold the phone at arm's length, breath in their full lung volume, and forcefully exhale at the phone microphone until the exhalation of the entire lung volume. They achieved a mean error of 5.1% on the prediction of standard lung function measures. The following work by Goel et al. [8] developed a system to reliably estimate (mean error of 6.2%) pulmonary function measures from a user's exhalation sound, through a call-in service on any phone using the standard voice telephony channel for transmitting the sound of spirometry effort. Another works called SpiroConfidence [10] attempted

to categorize valid spirometry test from invalid efforts using machine learning algorithms on blowing sound data captured using a smartphone microphone. Smartphone spirometry still remains highly susceptible to poorly performed efforts because of the variations in the effort.

Challenges with phone spirometry: First, the system needs to detect the right segment of the forced exhalation in spirometry from a continuous audio timeseries. Second, variation of distance between mouth and the device should be addressed. Third, the exhalation requires significant effort from the user and the lack of proper guidance can invalidate the results [10]. Therefore, estimating the quality of the blowing and guide the user to perform the blowing correctly is crucial to have a reliable measurement on mobile devices in uncontrolled field settings. Fourth, the model needs to handle device orientation variation to increase robustness and ease of use.

Novelty of our approach: Our approach handles the above mentioned challenges and move the concept of mobile sprimetry one step closer to the user space. We detect blowing exhalation sound to guide best quality spirometry maneuver on a smartphone. We evaluate our approach for both clinical and non-clinical dataset and show that our model can accurately assess the lung condition. We further evaluate the usability of our system with pulmonary patients by comparing it with a portable spirometer.

III. Study Design, Data Collection and Annotation

We have conducted three independent studies with three different cohorts of pulmonary patients in two different US locations in USA. We have collected data from 211 subjects, including chronic pulmonary patients and healthy subjects using Samsung Galaxy Note8 smartphones and Galaxy Gear Sport smartwatch. In the first two studies, we have collected several forced exhalation sounds from the participants from a one minute protocol. Participants were instructed to force their exhalation as long as possible and as quick as possible with multiple repetitions and comfortable breaks in between two efforts. The patient could stop doing the maneuvers if they feel uncomfortable at any moment. Based on their pulmonary condition and lung capacity, the duration of the forced exhalation sounds is expected to vary. In the third study, participants were given several doses of a bronchoprovocative medicine as a test to collect controlled variation of the pulmonary conditions. We have collected hospital-grade spirometry data along with forced exhalation sounds after each dose of the medication. This dataset is collected at BWH hospital, and all the above studies are approved by corresponding Institutional Review Boards (IRB). Participants were requested not to smoke or take medications several hours before the data collection. In all the studies, there were two smartphones - one for the patient to record the patient data, another for the study coordinator to label the start and end of each task (e.g., smartphone spirometry that includes multiple repetitions).

1) Study-I: Lab Study Using Connected Portable Spirometer: We have completed several iterations through pilot studies and mock-ups to make the study rigorous in terms of data quality and variability, and then, conducted it by partnering with a recruiting firm who provided access to pulmonary patients. We have collected data from 131 subjects. Among them, 91 are chronic patients (Male 41, Female 50), including 69 asthma patients, 9 COPD patients, and 13 with both conditions. The average age of the patients was 42.93 ± 19.49 years. Among the remaining 40 healthy subjects, 26 were male, and 14 were female. The ages of the subjects range from 14 to 82 years.

In this study, we used an FDA[2] approved, Bluetooth Low Energy (BLE) connected, portable spirometer to collect groundtruth FEV1, FVC, and FEV1/FVC ratio data. The spirometer was connected to a smartphone app, and the smartphone app could be able to assess the quality of the effort done by the participant. If the effort was not good enough, we asked the participants to perform the test again with greater effort. Moreover, we requested participants to perform the spirometry test on the spirometer up to three times so that we can choose the best effort from them as a groundtruth data. We also collect cough sounds, 'Aa....' vowel sound from

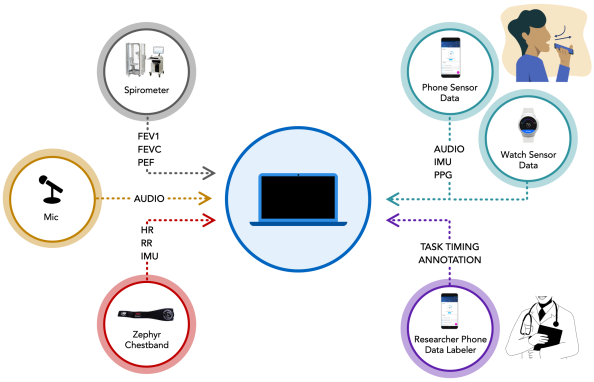


Fig. 2: Study setup for the mobile sensor data collection.

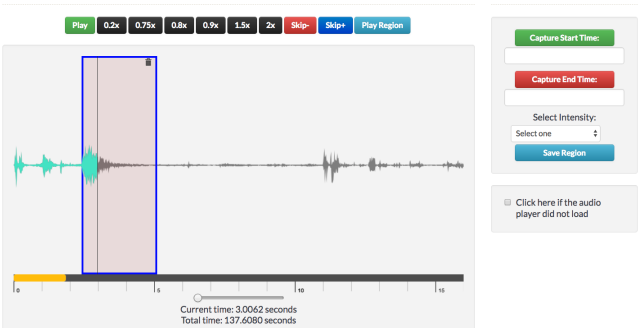


Fig. 3: Data annotation using FigureEight Inc crowdsourcing platform [17]. It shows the waveform visualization, and selection of audio segment that has a forced exhalation sound in a audio time series. It also shows that the platform gives ability to zoom-in and zoom-out the signal for further confirmation, and also to listen just the selected audio segment rather than listening the whole audio for confirmation. Moreover, the annotator could delete the annotation, if they find that the annotation was not correct after zoom-in or listening the selected segment.

[2]Food and Drug Administration

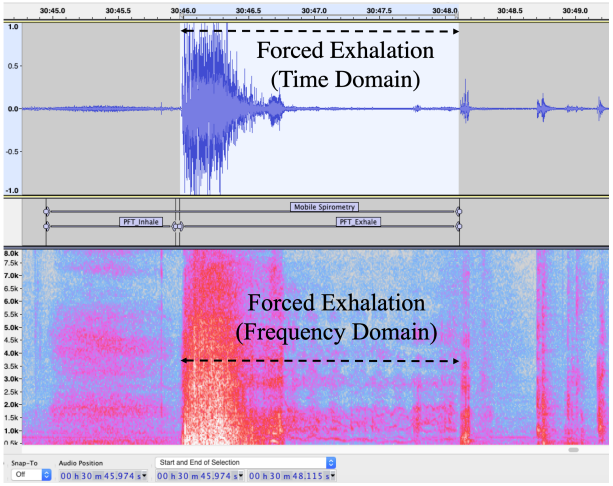


Fig. 4: Data annotation using Audacity [16] environment.

our participants which is out of the scope of this paper. The whole session is audio recorded including the instructions and exit interviews where we have collected usability questionnaire data.

2) Study-II: Clinical Study Using Hospital-grade Spirometer: We have conducted the second study in BWH hospital in Boston with an independent cohort of pulmonary patients. This study consisted of a total of 70 participants, including 25 asthma patients, 25 COPD patients, 10 chronic cough patients, and 10 healthy controls. Similar to the Study-I, we have collected smartphone phone spirometry, coughs, 'Aaa...' vowel sound data, hospital grade spirometry in a pulmonary function test lab, and usability questionnaire. The whole audio session is captured as a continuous audio timeseries data.

3) Study-III: Methacholine Challenge Test in Hospital: The methacholine challenge test is a standard bronchoprovocation test that helps diagnose asthma patient [14]. Methacholine is an inhaled drug that causes mild narrowing of the airways in the lungs, like asthma. At the beginning of the test, we have collected baseline spirometry test data using hospital-grade spirometer where the patient took a deep breath and blow their breath through the spirometer as hard as possible and as fast as possible, and we collected 'Aaa...' sound, voluntary coughs, speech, and spirometry (forced exhalation) on phone using Samsung Note8 smartphone. Then the participants inhaled methacholine medicine for five consecutive doses with a break of around 3-4 minutes. After each dose, we again collected the same set of data. If the patient's FEV1 value reduced below 20% after any of the doses (which means significant deterioration), the test was stopped. In the end, the patients were given bronchodilator medication to help their lung function recover. We have collected methacholine challenge test data to incorporate data with worsening lung condition and recovery from the deterioration into our model.

A. Data Annotation

Supervised machine learning models need high quality labeled data. We have two types of groundtruth data - (1) annotated forced exhalation audio segment and (2) lung function parameters (FEV1/FVC). We have collected the first type of ground truth from the hospital. Given the volume of the audio data collected in these studies, we have annotated the data in two ways - through (1) crowd-sourcing and (2) pulmonary researchers.

1) Crowd-sourced annotation: Traditionally, reliable pulmonary digital biomarkers and symptoms are annotated by expensive, expert medical practitioners, and health researchers. In the era of big data, crowd-sourcing for data annotation is proven to be a scalable inexpensive approach. We annotate approximately 44 hours (20 min x 131) of audio data using FigureEight Inc [17], a professional and secure crowd-sourcing platform specialized in annotating audio data. We conducted the annotation study with 12 randomly selected annotators in two phases. On average, each contributor had more than 9 months of experience with this platform. To obtain good quality annotations, the data annotators were further provided with extensive training materials on pulmonary sound annotation along with researcher-annotated sound samples.

Sound annotation interface: First, we segment the entire recording session into protocol task segments based on the timings annotated on the study coordinator's smartphone. We segmented those tasks into 1-minute chunks to improve the annotation quality and minimize annotator fatigue. We uploaded the anonymized and segmented data in a secured platform to crowd contributors with the Non-Disclosure Agreement (NDA).

Phase 1: Listening-based annotation: To help the crowd contributors understand and identify specific events of interest in an audio segment (event occurrence annotation), we provided the definition and examples corresponding to each sound event of interest (i.e., forced exhalation on the phone, cough, breathing, speech). The annotation interface shows a progress bar of the audio with time information. The interface enables annotators to play or pause the audio to listen, and mark the onset and the offset of each event by clicking on two respective buttons.

Phase 2: Visualization-based annotation: After the first phase and pulmonary researcher's expert review of the annotation reveals that only listening was not good enough for high quality pulmonary sound modeling as the annotation accuracy was low. Therefore, we introduced waveform visualization of the audio to improve the accuracy. It is because different pulmonary events (e.g., cough, spirometry) have different waveform patterns. Additionally, reviewers ranked the crowd contributors based on their accuracy and selected the most accurate contributors for the next iteration. We also updated the instructions based on the feedback collected from the crowd contributors. In a waveform visualization, the horizontal axis indicates the time and vertical axis indicates the audio amplitude of the sound. We also provided several examples of how to annotate, review, modify, and delete the onset and the offset of an annotated event. The annotation interface had the feature to zoom-in and zoom-out of the waveform for better pinpoint the onset and offset of pulmonary sounds (Figure 3).

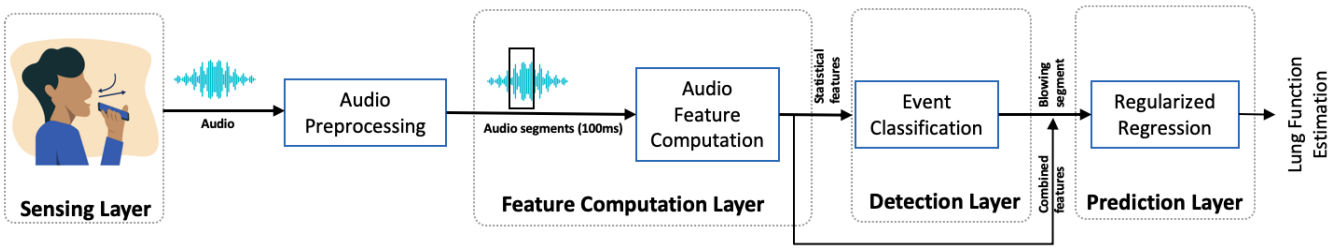


Fig. 5: Audio processing pipeline from sensing to forced exhalation detection to lung function estimation.

Quality of crowd-Sourced annotations: Researcher's review after the first phase of iteration (without waveform visualization) revealed that the crowd-sourced contributors annotated cough events with 72.2% accuracy, and forced exhalation on smartphones sounds with 90% accuracy. The speech annotation was reasonably accurate during the first phase.

Whereas the accuracy of pulmonary symptoms annotation improved significantly after the second phase of iteration. The cough annotation accuracy increased to 95.32%, and forced exhalation sound annotation accuracy becomes 91.07%. To establish that the learning effect does not confound our findings, we ran the annotation with the waveform visualization and updated instructions with a randomly chosen control group of four crowd contributors who did not have previous exposure to this dataset or annotating any pulmonary events. Researcher's review of their annotation shows that the control group annotated cough events with 92.27% accuracy and forced exhalation sounds with 92.7% accuracy. We demonstrated that the waveform visualization significantly helps to improve the quality of crowd-sourcing annotation for all the categories of pulmonary events.

Furthermore, all the forced exhalation sound annotations were visualized, listened, and verified to ensure that the ground-truth data are high quality and reliable.

2) Domain Experts' Annotation: Our team has domain expert researchers who have around 10 years of experience in analyzing pulmonary signals. They have further visualized the data annotated by the crowd annotators and adjusted the onset time of the forced exhalation events. This is very critical for the FEV1/FVC estimation model development since the flow from the initial one second is the most important portion of the whole forced exhalation event. If the start time of the forced exhalation is not annotated correctly, the resulting estimation will be highly unreliable.

We use crowd-sourcing for annotating the 131 subjects' data from the Study-I. Since the volume of the remaining data from Study-II and Study-III are smaller, our domain experts have performed highly reliable annotation using audacity audio annotation platform shown in Figure 4. It is to note that domain experts have visualized the audio signal as a waveform and a spectrogram in synchronization, and then listen the sound to confirm the start and the end of a forced exhalation event in the sound signal. The robust annotation and rigorous review protocol produce the labels for smartphone spirometry which are reliable to develop our model for the detection of forced exhalation sound and estimate lung function based on the detected sound.

IV. Model Development

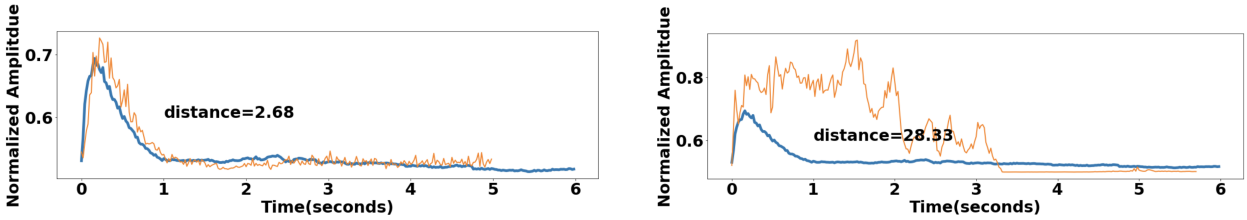
Audio processing pipeline (shown in Figure 5) to develop the model includes audio sensing, preprocessing, feature computation, forced exhalation detection and quality verification, and finally, lung function estimation. We describe each component below.

Audio sensing: we have collected audio data using Samsung Galaxy Note8 with 44100 Hz sampling frequency without any compression. The audio is captured in stereo mode. Therefore, sound events are captured by both channels. We have instructed the patient to hold the phone comfortably close to their mouth and blow their maximum breaths into the phone. We did not impose any restriction on how the patient would hold the phone, and also the distance between the phone and the mouth was based on participants comfort. It was intentionally done to mimic the real-world scenario when the patient would be doing the same test at home without any supervision. We observe that sometimes our participants were holding the phone horizontal to the ground and sometimes in a tilted position. Therefore, due to the variation of the orientation of the phone, one channel could capture high intensity sound then the other channel. For example, Figure 7 shows that the same forced exhale sound is captured by both channels and due to orientation variation, one channel captures the sound better than the other. Moreover, the intensity of the audio sensed by the smartphone microphone will also vary due to variation of distance between the mouth and the smartphone.

Audio pre-processing: To handle the variation of the distance between the mouth and the device, we have normalized the signal from 0 to 1 using Equation 1 which makes the forced exhale sound more prominent in the audio time series data. If X = $[x_1, x_2, ..., x_T]$ of duration T seconds of data from a particular channel, then we normalize X as

$$X_{normalized} = \frac{X - min(X)}{max(X) - min(X)} \quad (1)$$

Signal normalization is used to handle the variation of distance between the mouth and the smartphone that causes variation in amplitude in the sound signal. Furthermore, we have computed the root mean square (RMS) of the normalized signal from each channel and selected the channel that has the highest RMS value to handle the variation in device orientation.



(a) Low distance, high similarity

(b) High distance, low similarity

Fig. 6: Similarity between the template (blue) and the current forced exhalation (orange) from a smartphone spirometry effort.

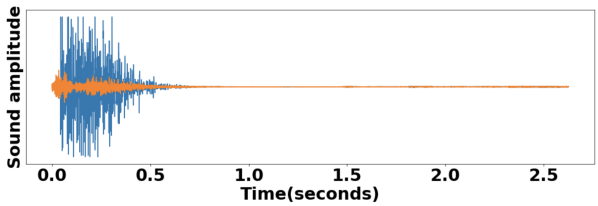


Fig. 7: Two channel audio captured during one forced exhalation maneuver for mobile spirometry. Root Mean Square (RMS) value for the blue channel is greater than the RMS value for the orange channel.

Based on American Thoracic Society and European Respiratory Society (ATS/ERS) [20] guideline, one forced exhalation for pulmonary function test should last between 5 to 6 seconds. Therefore, we segment the audio into 6 second overlapping window with a 300 millisecond shift. Since the forced exhalation sound has a unique shape in the time-domain audio waveform, we compute the envelope to capture this unique signature to detect its presence in the audio timeseries data. To ensure that the envelope is less affected by signal jitters, or other sharp noises, we compute the percentile based envelope rather than root mean square (RMS) based envelope. Therefore, we further segment the six seconds window into 10 milliseconds frames and compute 90th percentile over each frame to derive the envelope of a forced exhalation sound.

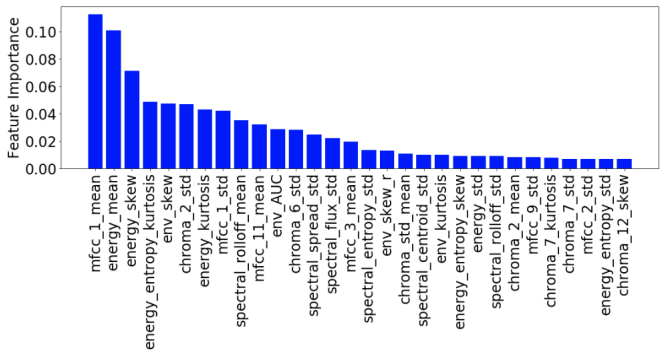


Fig. 8: Feature importance in detecting forced exhalation from other sound such as deep inhalation, cough, speech, and A-vowel sound.

Audio feature extraction: We compute both temporal and spectral features from each 6 second window by framing the window into 100 ms frames. Moreover, we also compute statistical features from the envelope signal. We compute statistical features such as mean, standard deviation, skewness, kurtosis of zero crossing rate, energy, spectral centroid, mel-frequency cepstral coefficients (MFCC), relative density, spectral flux, spectral rolloff, spectral entropy, spectral spread, and chroma features from each audio segment. Audio features other than envelopes are computed using PyAudioAnalysis [18] Python open-source library. Total 156 features are extracted from each window.

Forced exhalation event detection: We detect the presence of the forced exhalation in each window using machine learning classification algorithms. In the positive examples, we consider the annotated forced exhalation examples. In the negative examples, we include common pulmonary sounds such as coughs, deep inhalations, regular breathing sounds, 'Aaa...' vowel sound, other ambient noise (e.g., silence). We use 210 positive examples and 3953 negative examples to build our model. To reduce the skewness of the data, we downsample the negative class in a way so that it has similar number of samples, however, contains representative samples from each negative category. We split the dataset into train (80%) and validation set (20%). We further use 10-fold cross validation on the train dataset and use Random Forest ensemble classifier to detect forced exhalation event. Finally, the model performance is tested using left out validation dataset.

Effort quality assessment: Once we detect a window of 6-second audio as a forced exhalation segment, we further pinpoint the start of the exhalation by removing the initial silence based on an empirically learned energy threshold. We then assess the quality of the forced exhalation effort based on the American Thoracic Society guidelines [20]. It is to note that the ATS/ERS guidelines are mostly related to the traditional spirometer where the patient needs to insert a tube in their mouth and blow their breaths into the spirometer. We observe that the audio of the forced exhalation is mostly audible in first 2-3 seconds out of the 6 second effort. Therefore, the most important quality parameter from ATS/ERS that is applicable for smartphone spirometry is the Time To Peak Flow (TTPF). We compute the TTPF and if it is below the ATS/ERS threshold (less than 300 milliseconds), we consider it as a poor effort and discard from the assessment.

To ensure that we detect the high quality effort and provide appropriate feedback to the patient who is using the smartphone spirometry at home without any clinician supervision, we further analyze the shape of the envelope of a forced

expiratory effort sound. We develop a shape based time series data modeling approach to detect a template of the forced exhalation sounds from the rigorously annotated sound segments. We use percentile based envelope approach described in audio preprocessing subsection to generate individual envelope. We compute the mean of the individual envelopes from the training data to generate the template envelope for high fidelity forced expiratory effort sounds (Figure 9). We compute the TTPF for the template envelope for the sanity check. We find that it is less than 300 milliseconds which ensures that the envelope template also meets the ATS/ERS criterion.

We propose the similarity between the template and the individual envelope as a metric for smartphone spirometry. We compute the similarity as the absolute distance between the template (which is the expected shape) and the envelope from individual forced expiratory sound segment as described in the equations below. First, we describe the construction of the template:

$$T[n] = \frac{1}{K}\sum_{k=1}^{K} Env_k[n], \quad \forall\, n \in [1, N] \tag{2}$$

where $T[n]$ is the n^{th} sample of the envelope template T, $Env_k[n]$ is the n^{th} sample of the k^{th} envelope Env_k in the training set, K is the total number of envelopes that have an available sample at the n^{th} position, and N is the total number of samples in the longest envelope. The distance metric is then computed as the absolute difference between each of the data points of the current envelope Env_c and the template envelope T as the equation given below:

$$D = \sum_{i=1}^{M} ||Env_c[i] - T[i]|| \tag{3}$$

where M is the total number of samples in the current envelope Env_c, and it is assumed that $M <= N$, where N is the total number of samples in the template T.

We compare the similarity between the template and the domain expert annotated forced exhalation sounds. We observe that the higher the distance, the lower the quality of the effort (Figure 6). From our dataset, we compute the similarity metric for all the smartphone spirometry efforts and plot the distribution of the distance. We observe that similarity distance more than 15 divides the high quality efforts from the poor efforts (Figure 10).

Regularized regression to estimate lung function: We predict lung obstruction parameter called FEV1/FVC ratio. Since the forced expiratory flow during the initial one second is critical for the lung obstruction estimation, we compute the same audio features from the first one second of the window to enforce importance of the initial one second in the model. Therefore, we have 312 features for lung obstruction estimation. We then follow a hierarchical feature selection approach to select the best features among them to estimate lung obstruction. First, we apply linear regression with L1 regularization which selects

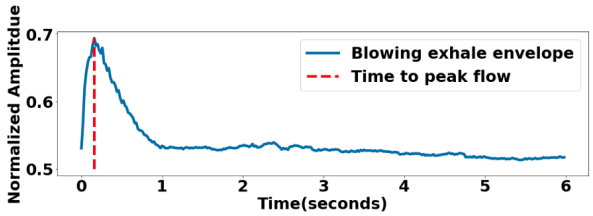


Fig. 9: Template from high quality blowing exhalation on a smartphone microphone while doing the spirometry test on the phone.

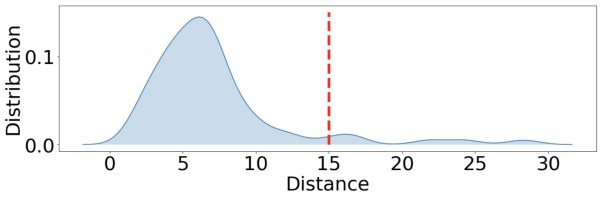


Fig. 10: Distribution of the distance between the template and the smartphone spirometry efforts. It seems that cut off threshold divides the high quality efforts from the poor efforts.

185 features. Then, we compute correlation (Pearson's r) with the target variable (FEV1/FVC ratio from spirometers). We select the features that are significantly correlated ($p < 0.05$) with the target variable. Then, the features set reduces to 45. Finally, we use this feature set to train a linear regression with ridge regularization.

We split the dataset into train (90%) and validation (10%) set. We train a L2 regularized linear regression with 10-fold cross validation and test the model on the validation dataset. We do this split randomly 100 times.

V. Results

Forced exhalation detection performance: As described in the previous section, we develop a binary classifier that detects forced exhalation effort sounds from other sounds. Using Random Forest classifier, we can achieve F1-score of 96.74%+/-1.84%. We find that the most discriminating features are the mean of first mel-frequency cepstral coefficient, energy, kurtosis of energy entropy, skewness of envelope, standard deviation of the second chroma feature are the top five important features in identifying a forced exhalation from other sounds in a audio timeseris data. Figure 8 shows the feature importance from high to low for our model to detect forced exhalation. It is to note that MFCC represents spectral envelope of a sound signal. We observe that energy features of the forced exhalation effort sound, spectral envelope feature, and the percentile envelope features are having more discriminatory power to identify a smartphone spirometry event.

Quality assessment performance: In addition to following the time-to-peak-flow (TTPF) recommended by ATS/ERS standard, we develop that distance based quality metric. We observe from the distribution of the distance metric in our dataset that distance cut-off 15 can identify around 93.3% of the high fidelity smartphone spirometry effort from the poor efforts (Figure 10). It demonstrates the strength of our distance

TABLE I: Top ranked features after doing LASSO regularization and correlation based selection to estimate FEV1/FVC ratio. Feature names with _1s indicate that those features are computed from initial one second of the window. Other features are computed from the whole window.

Rank	Feature Name	Rank	Feature Name
1	Chroma5_kurtosis	13	MFCC7_skewness
2	Exhalation_duration	14	Chroma_std_mean_1s
3	MFCC5_skewness_1s	15	MFCC1_skewness
4	Chroma6_kurtosis	16	Chroma_11_kurtosis
5	Spectral_density[1.6k-1.7k]_1s	17	Chroma_11_skewness
6	Spectral_density[1.6k-1.7k]	18	Chroma8_skew_1s
7	Chroma6_skewness	19	Spectral_spread_skewness
8	Envelope_slope_1s	20	Envelope_area
9	Chroma10_kurtosis	21	MFCC4_skewness_1s
10	Chroma10_skewness	22	MFCC4_kurtosis
11	Chroma3_kurtosis	23	Spectral_entropy_skewness
12	Chroma_std_skewness	24	Spectral_spread_kurtosis_1s

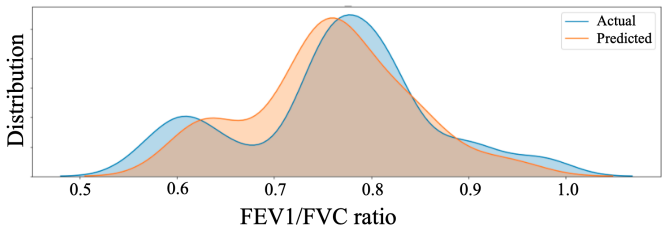


Fig. 11: Prediction performance of FEV1/FVC ratio using Linear Regression with Ridge regularization. Mean Absolute Error (MAE) for this model is 7.57%.

metric to distinguish high fidelity forced exhalations from poor efforts. Future works can explore how this accuracy can further be improved by this distance metric in conjunction with other envelope parameters such as mean, variance, and correlation between the envelope and the template envelope.

Lung obstruction estimation performance: We trained a linear regression model with L2 regularization based on 45 best features and predict lung obstruction measures called FEV1/FVC ratio (a continuous value between 0 and 1). We observe that our cross validated model can predict FEV1/FVC ratio with mean absolute error (MAE) of 7.57%. Figure 11 shows that the distribution of the actual FEV1/FVC ratio values in test dataset and predicted FEV1/FVC ratio values are having quite similar distribution. As we mentioned in the model development section, the lung obstruction prediction model includes features from both initial one second and the whole window of a forced exhale effort. We observe that chroma5 kurtosis, exhalation duration, MFCC5 skewness from initial one second, chroma6 kurtosis, spectral density between 1600Hz-1700Hz band from initial one second, spectral density between 1600Hz-1700Hz band from the whole window, chroma6 kurtosis, envelope slope from initial one second, chroma10 kurtosis and skewness are the top 10 features. Several of the most important features selected by LASSO and Pearson's r based feature selection method are shown in Table I.

Patient comfort with smartphone spirometry: In Study-I and Study-II, we asked the participants on their preferences between mobile phone spirometry and connected portable spirometry, and between smartphone spirometry and hospital-grade spirometry. We find that 42.18% participants in Study-I preferred phone spirometry whereas 39.06% preferred blowing on the portable spirometer. The reasons of choosing phone spirometry were the availability of the devices, easy to use, doesn't need anything to insert into mouth, among several others. The reasons of choose portable spirometry were the tube that helps focus the effort and the blow for the patients compare to the phone spirometry. We see a bigger difference in the Study-II where 72.41% participants preferred phone spirometry over the hospital grade spirometry (13.79%). Perhaps, it is because the hospital-grade spirometry was a big machine, need to sit in a special housing, and overall, it was more cumbersome and effortful experience for the patients compare to the mobile spirometry.

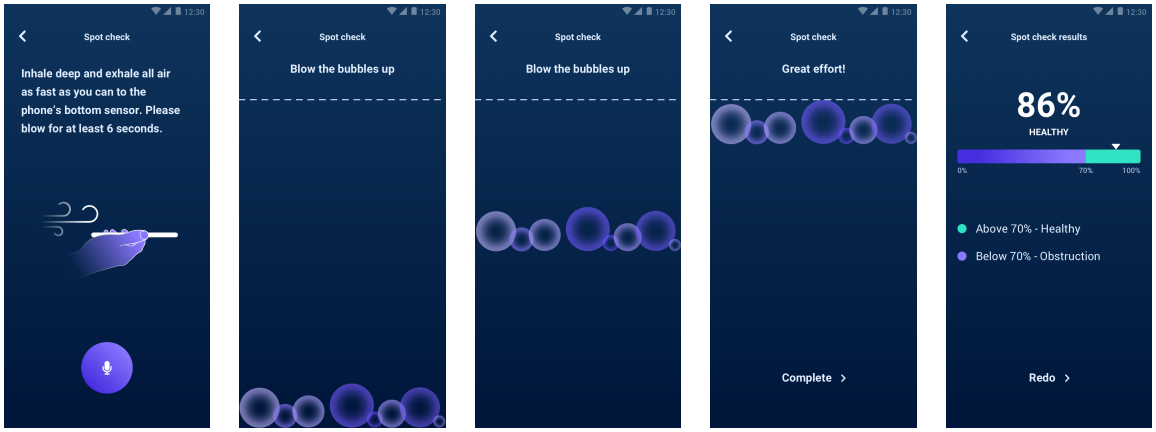
VI. Application and User Behaviors

We have implemented the model on device using Samsung Note8 smartphone. This is an on-demand app where the user can press the start button and perform the spirometry on-device. The data is not sent to the server in order to preserve user privacy. To understand the user behavior, we give the app to 10 subjects to follow the instructions on the phone screen and try a smartphone spirometry test. The instructions are the following: "Inhale deep and exhale all air as fast as you can to the phones bottom sensor. Please blow for at least 6 seconds". We didn't provide any further encouragement or guidance as we want to understand actual behavior of a new user. Each participant forcefully exhales their breaths three times on the phone to measure their lung obstruction. We find that the time gap between the start button and the actual blowing affects the accuracy of the model. From this experiment, we find that time gap between the start button press and the actual blowing is 0.95+/-0.56 seconds. It further supports our motivation of detecting and pinpointing the forced exhalation effort sound which is crucial for field deployment. It is because forced expiratory volume in the initial one second is the most important lung function. Therefore, first one second cannot be just noise or silent sound for high fidelity pulmonary function assessment.

VII. Discussion, Limitation, and Future Works

Ensuring the quality of the patient effort during the spirometry test is one of the biggest challenges for pulmonary function test. Our model error is around 7.57%. Usually, 15-20% decline in lung function is considered the significant deterioration or pulmonary exacerbation [19]. In future, we'll incorporate our quality metric to model the lung function change from our rigorously collected clinical dataset to see whether our model can predict significant lung deterioration.

Since smartphone spirometry at home, in free-living condition, is still a promising novel direction, there can be a lot of future studies to understand the performance of the system at home. Forced exhalation detection and effort estimation on smartphone can help quantify patient engagement and compliance in those field studies. For example, if the participants is asked to do smartphone spirometry once a day, the study coordinator may want to know whether the participants is



(a) Instructions (b) Blowing Start (c) Blowing Stage (d) Blowing Complete (e) Estimated Results

Fig. 12: Design of our intuitive interface for patient encouragement to achieve high quality effort on smartphones. This figure shows that stages of a smartphone spirometry test from sensing to lung condition estimation.

indeed performing the test on the phone to generate high quality data for model development.

Our studies were conducted in lab and hospital environment. Although there are noises from other patients, nurses, and the machines around the patients, the noise at home can be different from the noise in our data. Therefore, it warrants studies at home to understand the performance of our algorithm in free-living conditions.

VIII. Conclusion

Smartphone spirometry has a great potential to make pulmonary assessment and tracking available anywhere, anytime. However, the lack of methods to reliably assess patient's spirometry effort on smartphone at home is a big challenge. This paper takes an important step towards moving the smartphone spirometry from feasibility to the consumer space. Based gold-standard spirometers and rigorous data annotation, we present models to detect forced exhalation efforts on smartphones, assess the quality of the effort, and estimate the lung obstruction. We show that our model can detect forced exhalation with 96.74% F1-score and estimate lung obstruction with less than 8% mean absolute error. We propose explainable metrics to assess quality of smartphone spirometry effort which is informed by American Thoracic Society and European Respiratory Society guidelines. We believe that our approach makes a strong impact towards taking the smartphone spirometry into patient's home.

Acknowledgements

We want to thank Matthieu Chaminade and Philip Park for the app interface design, Hujun Cui for the android framework development, Leonardo Jimenez Rodriguez for preparing the back-end servers, crowd contributors for data annotation, Nazir Saleheen for data analysis feedback, Daniel McCaffrey for managing the multi-disciplinary project, Keiko Kurita for the review and valuable feedback, Alex Gao for his overall feedback in the project and the presentation.

References

[1] H. Ritchie, and M. Roser, "Causes of Death. https://ourworldindata.org/causes-of-death,", in 2019.
[2] World Health Organization, "Chronic obstructive pulmonary disease (COPD). https://www.who.int/en/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)," in 2019.
[3] Masekela, Refiloe and Zurba, Lindsay and Gray, Diane, "Dealing with Access to Spirometry in Africa: A Commentary on Challenges and Solutions," in Journal of Environmental Research and Public Health, 2018.
[4] Juen, Joshua, Qian Cheng, Valentin Prieto-Centurion, Jerry A. Krishnan, and Bruce Schatz. "Health monitors for chronic disease by gait analysis with mobile phones." in Telemedicine and e-Health, 2014.
[5] Juen, Joshua, Qian Cheng, and Bruce Schatz. "A natural walking monitor for pulmonary patients using mobile phones." in IEEE Journal of biomedical and health informatics, 2015.
[6] Cheng, Qian, Joshua Juen, Shashi Bellam, Nicholas Fulara, Deanna Close, Jonathan C. Silverstein, and Bruce Schatz. "Predicting pulmonary function from phone sensors." in Telemedicine and e-Health, 2017.
[7] Infante, Christian, Daniel Chamberlain, R. Fletcher, Y. Thorat, and Rahul Kodgule. "Use of cough sounds for diagnosis and screening of pulmonary disease." in IEEE Global Humanitarian Technology Conference, 2017.
[8] Goel, Mayank and Saba, Elliot and Stiber, Maia and Whitmire, Eric and Fromm, Josh and Larson, Eric C. and Borriello, Gaetano and Patel, Shwetak N., "SpiroCall: Measuring Lung Function over a Phone Call," in ACM SigCHI 2016.
[9] A. Mariakakis, and E. Wang, and S. Patel, and M. Goel, "Challenges in realizing smartphone-based health sensing," in IEEE Pervasive Computing, 2019.
[10] V. Viswanath, and J. Garrison, and S. Patel, "SpiroConfidence: Determining the Validity of Smartphone Based Spirometry Using Machine Learning," IEEE EMBC, 2018.
[11] U. Melia, and F. Burgos, and M. Vallverdú, and F. Velickovski, and M. Lluch-Ariet, and J. Roca, and P. Caminal, "Algorithm for automatic forced spirometry quality assessment: technological developments," PloS one, 2014.
[12] E. J. Sakka, and P. Aggelidis, and M. Psimarnou, "Mobispiro: A novel spirometer," in Springer Conference on Medical and Biological Engineering and Computing, 2010.
[13] E. C. Larson, and M. Goel, and G. Boriello, and S. Heltshe, and M. Rosenfeld, and S. Patel, "SpiroSmart: using a microphone to measure lung function on a mobile phone," in ACM UbiComp 2012.
[14] R. O. Crapo, "Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999," American Journal of Respiratory and Critical Care Medicine 2000.

[15] Zhou, Ping and Yang, Liu and Huang, Yao-Xiong, "A smart phone based handheld wireless spirometer with functions and precision comparable to laboratory spirometers," in MDPI Sensors 2019.
[16] Audacity, "https://www.audacityteam.org", in 2019
[17] Figure Eight Inc., https://www.figure-eight.com/, in 2019.
[18] G. Theodoros, "pyaudioanalysis: An open-source python library for audio signal analysis," in PlOS One 2015.
[19] J. M. Bhatt, Jayesh M. "Treatment of pulmonary exacerbations in cystic fibrosis." in European Respiratory Review, 2013.
[20] M.R. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C.P.M. van der Grinten, P. Gustafsson, R. Jensen, D.C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O.F. Pedersen, R. Pellegrino, G. Viegi and J. Wanger, "Standardisation of spirometry," in European Respiratory Journal 2005.