Towards Passive Assessment of Pulmonary Function from Natural Speech Recorded Using a Mobile Phone

Keum San Chun*, Viswam Nathan*, Korosh Vatanparvar*, Ebrahim Nemati*, Md Mahbubur Rahman*, Erin Blackstock†, Jilong Kuang*
*Digital Health Lab, Samsung Research America, Mountain View, USA
†Pulmonary and Critical Care Division, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
gmountk@gmail.com

***Abstract*—Chronic obstructive pulmonary disease (COPD) and asthma are the most common respiratory diseases that impact millions of people worldwide annually. With advances in mobile computing and machine learning techniques, there has been increased interest in using mobile devices to monitor pulmonary diseases. Nevertheless, the current state-of-the-art technology requires active involvement and high-effort input from the users, impeding continuous monitoring of pulmonary conditions. In this work, two algorithms are proposed for passive assessment of pulmonary condition: one for detection of obstructive pulmonary disease and the other for estimation of the pulmonary function in terms of FEV_1/FVC ratio, which is an established clinical metric. The algorithms were developed and validated using the data sets from two studies: research study (healthy=40, pathological=91) and in-clinic study (healthy=10, pathological=60). From the cross-study validation where a classifier was trained on the research data set and tested on the in-clinic data set, the detection accuracy of the pathological class was obtained as 73.7% and the F1 score was 84.5% (87.2% precision and 82.0% recall). In our regression analysis, the FEV_1/FVC ratio was predicted with a mean absolute error of 8.6%. Our analysis shows promising results and this work presents a meaningful milestone towards the passive assessment of pulmonary functions from spontaneous speech collected from a mobile phone.**

***Index Terms*—Mobile Computing, Pulmonary Assessment, COPD, Asthma, Natural Speech, Inspiratory Sound**

I. Introduction

Chronic obstructive pulmonary disease (COPD) and asthma are the most common respiratory diseases affecting greater than 500 million people worldwide [1]. Among the deaths caused by non-communicable diseases(NDS), respiratory diseases were reported as the fourth leading cause with 4 million deaths in 2012 [2], and this number is projected to continuously increase [3]. The high economic burden associated with this increasing prevalence is another problem. According to Nurmagambetov *et al.*, the total cost of asthma in the United States was estimated to be $81.9 billion in 2013 [4], and the total healthcare cost associated with COPD in the United States was estimated to be $32.1 billion in 2010 and is expected to increase to $49 billion by 2020 [5].

Asthma and COPD are classified into four stages based on the severity of the symptoms [6], [7]. There is strong evidence that early diagnosis plays an important role in delaying the progression of symptoms to more severe stages, leading to a better prognosis [8]. However, the early symptoms are often too subtle to be sensed by the patients, and thus, patients often miss the optimal window for treatment [9]. In case of asthma exacerbations, experimental testing with human rhinovirus (HRV) inoculation indicates that peak upper respiratory symptoms appear within 2-4 days from infection [10]. With earlier intervention, it could be possible to prevent exacerbations requiring ER visits and hospitalizations [9]. However, to realize early intervention, more frequent assessment of physiological conditions is a prerequisite. Given today's hospital-centric paradigm of health care, this requirement of frequent assessment of pulmonary conditions presents a challenge since people tend to visit hospitals only when they get sick.

With advances in mobile computing technology, there has been increased interest in leveraging mobile and wearable devices for monitoring pulmonary conditions. Utilizing mobile devices presents many advantages compared with traditional, on-site pulmonary examinations. One advantage is portability; mobile devices can be carried around and stay closely coupled with the individual. Thus, the portability aspect alleviates the limitation imposed on the physical location (e.g. hospitals) of pulmonary assessment. Also, since mobiles devices are already widely used, a large number of people can readily utilize the new developments without having to purchase additional devices. Nonetheless, there are many challenges that need to be overcome to reliably use mobile-based pulmonary assessment approaches. One of the greatest challenges would be the reliability and accuracy of the mobile-based approach. Spirometry, practiced in hospitals for pulmonary assessment, is an accurate and established medical examination whereas pulmonary assessment using mobile devices needs to be validated and proven to be accurate.

In this paper, we present novel features and algorithms in an attempt to extend the boundaries of mobile-based pulmonary assessment technology. In our approach, we leverage spontaneous speech sound collected from a smartphone to 1) detect pulmonary diseases and 2) predict pulmonary function

in terms of FEV_1/FVC ratio, an established spirometry parameter frequently employed in clinical settings. FEV_1 represents the forced expiratory volume in 1 second, and FVC is the forced vital capacity; the ratio of the two is a measure of pulmonary obstruction. Being able to leverage the daily activity of natural speech for low-effort, passive assessment of lung function implies a significant leap towards ubiquitous monitoring of pulmonary functions. The contributions of this work are summarized below:

- The development of novel features from inspiratory sounds
- The development of a detection algorithm that detects obstructive pulmonary disease from natural speech
- The development of a regression algorithm that predicts the user's FEV1-to-FVC ratio from natural speech.
- Validation of the approaches using two data sets that are independently collected from different environments

II. Related Works

A. Electronic Diaries

Johnston *et al.* investigated the possibility of using a BlackBerry smartphone for tracking COPD patients and detecting COPD exacerbations [11]. Over a 2-year period (from December 2007 until April 2009), 79 COPD patients with severity ranging from GOLD stage I to IV, were alerted to complete a daily diary with questions regarding their current symptoms. Their answers were used to assess the progression of their symptoms and apply appropriate interventions to the patients. Johnston et al. reported that near complete detection of COPD exacerbations was possible using electronic diaries.

Chan *et al.* conducted a large-scale clinical observational study where they monitored 7,593 asthma patients from the United States using a smartphone application called Asthma Health Application (AHA) for a 6-month study period [12]. The participants were asked to answer daily asthma surveys to record their symptoms, presumed triggers, and compliance for medication intake [12]. The contribution of the work by Chan *et al.* was that through electronic diaries, they were able to identify environmental risk factors or triggers associated with asthma.

B. Mobile Spirometry

Many mobile spirometers have been developed and are commercially available today [13]–[16]. However, the commercially available spirometers require dedicated hardware that needs to be purchased and carried. As such, there have been attempts to leverage the smartphone's computing power to enable mobile spirometry without requiring any additional hardware. In 2012, Larson *et al.* presented SpiroSmart, a mobile application that evaluates pulmonary function via similar assessment procedure as the clinical spirometry [17]. SpiroSmart utilizes the sound captured by the microphone to estimate the flow exiting a patient's mouth. SpiroSmart was evaluated on 52 participants, and the mean absolute error (MAE) to clinical spirometer was 5.1% [17]. In 2016, Goel *et al.* presented SpiroCall, an algorithm that works through the GSM network to enable spirometry over the call [18]. Using SpiroCall, anyone with access to GSM network can have their pulmonary function tested over a phone call.

C. Continuous Monitoring

While electronic diaries and mobile spirometry show promising results, the approaches necessitate active involvement of the users. For electronic diaries, the users need to manually input their response every day, and this approach is prone to recall bias. Mobile spirometers require users to master performing the spirometry maneuver without the oversight of a respiratory technician. Such requirements for active input from the users make continuous assessment of pulmonary function burdensome for the user and the individual assessments susceptible to errors.

A number of different studies have investigated the use of sound data, which contains rich information about pulmonary symptoms and allows passive data collection. Many studies focused on common pulmonary symptoms such as cough, shortness of breath, and wheeze [19]. Anderson *et al.* explored the possibility of detecting wheeze, one of the common symptoms in asthma, using sound analysis from mobile devices [20]. Shaharum *et al.* analyzed wheeze sound to classify the severity of asthmatic symptoms [21]. Larson *et al.* presented an algorithm that detects coughs from a stream of audio collected from a mobile phone [22].

Despite the significant body of work dedicated to leveraging sound data as exemplified in the aforementioned studies, most of the previous work mainly targeted the symptoms. Thus, there remains the question on whether or not the speech data can be utilized for detecting pulmonary disease and assessing pulmonary function for these patients. To the best of our knowledge, only a few studies have explored the possibility of identifying patients with pulmonary disease from speech data. Nathan *et al.* presented a work on passive detection of obstructive pulmonary diseases such as COPD and asthma from scripted speech [23]. In the study, over 90 pulmonary patients and 40 healthy subjects were recruited, and their pulmonary conditions were assessed with spirometry. Then, a set of passages was provided to each participant, and the reading sound was recorded using a smartphone. The recorded audio data was analyzed to develop a classification algorithm that can detect patients with pulmonary disease [23]. Nathan *et al.* reported the overall accuracy of 65% and F1-score of 62% (61% precision and 65% recall) [23].

III. Background and Motivation

The gold standard approach for pulmonary function test (PFT) today is spirometry. Spirometry is performed by having a person pinch his or her nose, inhale to maximum lung volume, and then forcefully exhale. Despite the reliability of the approach, spirometry is rarely accessible to individuals outside of clinical settings due to the requirement for a specialized device. Moreover, performing the spirometry maneuver correctly requires training, and in most in-clinic settings there is a clinical professional present at the time of spirometry

to encourage a proper effort. In addition, as a high-effort examination requires the user to repeatedly forcefully exhale, spirometry is difficult to be used for continuous monitoring of pulmonary function. As mobile computing technologies and machine learning advance, an extensive body of work has been dedicated to pulmonary monitoring as exemplified in the previous section. Despite the efforts, the current state-of-the-art technology still requires high-effort activity (e.g. forceful blowing) from the user.

Speech is the generation of meaningful sounds that are modulated by the vocal fold movements. The pressure generated by the diaphragm forces the air out through the airways, and as the air passes through the airways and vocal folds, the sound is created. Due to this sound generation mechanism, the speech signal contains rich information about the airways as well as vocal folds. With this as the basis, many studies explored the possibilities of leveraging speech analysis to assess the severity of pulmonary diseases, and promising results have been reported [19]–[21]. Nevertheless, speech alone has not been extensively explored in the context of detection of pulmonary diseases and assessing pulmonary function.

According to Klemmer, people spend approximately 50-80% of their work time communicating, and two-thirds of the communication time is spent talking [24]. If this common activity can be leveraged for low-effort, passive assessment of pulmonary functions, continuous assessment of people's pulmonary functions would be possible. This notion of leveraging speech for pulmonary disease detection and pulmonary function assessment motivated our research.

IV. Approach

In this study, our goals were to develop algorithms that can detect obstructive pulmonary diseases and assess pulmonary functions from natural speech. To this end, the speech data were processed through the 4-step signal processing pipeline shown in Figure 1. The first step was the frame generation step using a sliding window method. In this step, a window 20 milliseconds in size was moved across the audio data without overlap, resulting in a stack of frames. In the second step, the stack of frames was input to the speech detection block, which identified the speech and pause frames using the Long-Term Spectral Divergence (LTSD) algorithm. The third step extracted features using the pause and speech frames. For clarification, each participant's audio data generated a single set of features at the end of the third step. In the fourth step, the generated features from multiple participants were used to build and test a classification and regression model. A more detailed description of the approach is provided below.

A. Voice Activity Detection

Speech data is a mixture of voiced segments and pause segments. This is because people take pauses to inhale between each voiced segment. Prior to extracting features from the audio file, we identified voiced segments and pause segments from the speech audio data using the LTSD-based algorithm proposed by Ramirez *et al.* [25] as the basis. This algorithm looks at the long-term divergence between the noise spectra and potential voice spectra. We chose the LTSD-based voice activity detection (VAD) algorithm because this approach was rather simple to implement and robust to background noise [25]. The VAD algorithm used in this study starts from defining the order, which describes how many frames to account for before and after the current frame in computation of LTSD. In this study, the order was set as 10. After defining the order, we compute the long-term spectral envelope (LTSE). The LTSE was obtained by getting the maximum spectral amplitude from the frames within the range defined by the order. Hence, in our example case, LTSE will be the spectral envelope of 21 consecutive frames (10 frames before and after the current frame). Then, LTSD was computed according to the following formula:

$$LTSD_k(l) = 10\log_{10}\left(\frac{1}{NFFT}\sum_{i=0}^{NFFT-1}\frac{LTSE_k^2(i)}{N^2(i)}\right)$$

In the formula, k indicates the frame number (*i.e.,* k-th frame). $NFFT$ is the number of frequency bands and N is the average noise spectrum magnitude. In our VAD algorithm, we assumed that the lowest 5% of given audio recording was comprised purely silence or any background noise, and not the primary user's speech data. With this assumption, we used these segments for constructing the noise profile.

B. Inspiration Filtering

The act of vocalization process involves continuous exhalation. To compensate for the loss of air during the voice activity, there are often sharp inspiration periods right before or after long vocalization periods. In order to extract the inspiration segments, we further processed the pause segments. Inspiration sound is generally independent of the vocal fold movements, but rather more influenced by the obstruction in the airway and the rapidness of inspiration [26]. While slow inspiration tends to make a quiet sound, rapid inspiration tends to generate relatively louder, high energy signal. Based on this observation, inspiration frames were identified via two steps. In the first step, the potential inspiration frames were identified by computing the energy in the given pause segment and selecting the frames whose energy falls in between the 60^{th} and 95^{th} percentiles. In the second step, the selected frames are clustered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The extracted inspiration frames were input to the feature extraction block along with the voiced frames.

C. Feature Extraction

Two types of features were employed in our work: acoustic features and pulmonary features. The acoustic features are included to account for dysphonia, a comorbidity frequently observed in pulmonary patients [27], [28], and the pulmonary features are incorporated to indirectly extract information about the pulmonary obstruction. Acoustic features include shimmer and jitter. Pulmonary features include pause frequency, vocalization to inhalation ratio, average phonation

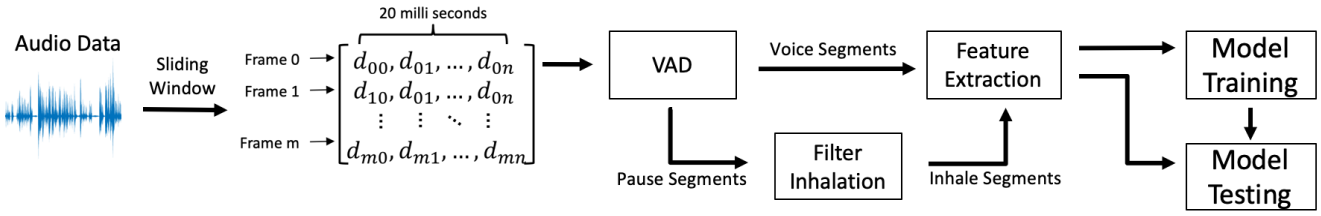

Fig. 1. Overview of the signal processing approach. Model testing consists of leave-one-participant-out (LOPO) or 10 iterations of 10-fold cross-validation with data shuffling in each iteration.

time, and inspiration sound energy. Initially, we explored other features such as harmonic-to-noise ratio (HNR), mean pause time, variance in inspiratory periods, etc.; we narrowed down to the set of 7 features which resulted in the best classification performance. A more detailed description of the feature selection is provided in Section VII-E. A detailed description of the selected features follows in this section.

1) Shimmer (apq3): Shimmer is a measure of cycle-to-cycle amplitude perturbation of the dominant frequencies. In healthy individuals, shimmer value is smaller compared with the shimmer values from individuals with dysphonia. A high shimmer value may be an indication that one's pulmonary symptom is getting worse. Shimmer can be calculated in several different forms such as three-point amplitude perturbation quotient (apq3), five-point (apq5), or eleven-point (apq11). In our analysis, we used apq3 to compute the shimmer. The specific shimmer measure we used was obtained using the following equation expressed in percentage:

$$\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100\%$$

In the equation, N indicates the number of periods of the fundamental waveform in the given vowel sound, and A_i indicates the amplitude of the i-th peak of the fundamental waveform.

2) Absolute Jitter: Jitter is another acoustic feature that measures the cycle-to-cycle frequency perturbation of the dominant frequencies. As with shimmer, higher jitter values may indicate greater possibility of dysphonia in voice and worsening of pulmonary symptoms. The absolute jitter is computed by taking the average of variations in periods between every two consecutive peaks of the fundamental waveforms. The absolute jitter was computed according to the following equation and was expressed in microseconds:

$$\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|$$

In the equation, N indicates the number of periods of the fundamental waveform, and T_i indicates the duration of the i-th period of the fundamental waveform in the given vowel sound.

3) Relative Jitter: The relative jitter is the average variations in periods between every two consecutive peaks of the fundamental waveforms, relative to the overall average period. The relative jitter was computed according to the following equation and expressed as a percentage:

$$\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100\%$$

As with the absolute jitter, N is the number of periods of the fundamental waveform, and T_i is the duration of the i-th period of the fundamental waveform in the given vowel sound.

4) Pause Frequency: For patients with obstructive pulmonary disease, breathing is made more challenging due to the increased airway resistance [29]. They frequently note increased breathlessness and may have increased respiratory rates when their pulmonary symptoms are exacerbated. Based on this reasoning, we captured the information related to the shortness of breath by computing the average number of pauses per minute in each user's speech.

5) Vocalization to Inhalation Ratio: As the voice is generated, a small amount of air is expelled continuously from the lung through the airway (Figure 2A). Due to this steady loss of air and pressure inside the lung, quick inspiration is required during speech to replenish the air in the lung (See Figure 2A). Since pulmonary disease can affect the ability to speak for long periods and the amount of air inhaled at the end of the speaking period, we considered the inhalation duration during the pause for breath relative to the immediately preceding vocalization. This feature is hereby referred to vocalization to inhalation ratio. We built a simple model for lung volume change during speech (See Figure 2B). In this model, we assumed that the rate at which lung volume decreases during the vocal period was a constant, denoted φ, and approximated the relative inhalation duration using the durations of the corresponding vocalization and inhalation sounds. (See Figure 2B).

6) Average Phonation Time: Vocalization may exert additional burden in patients with dyspnea (also known as shortness of breath), as it requires higher than usual expiratory

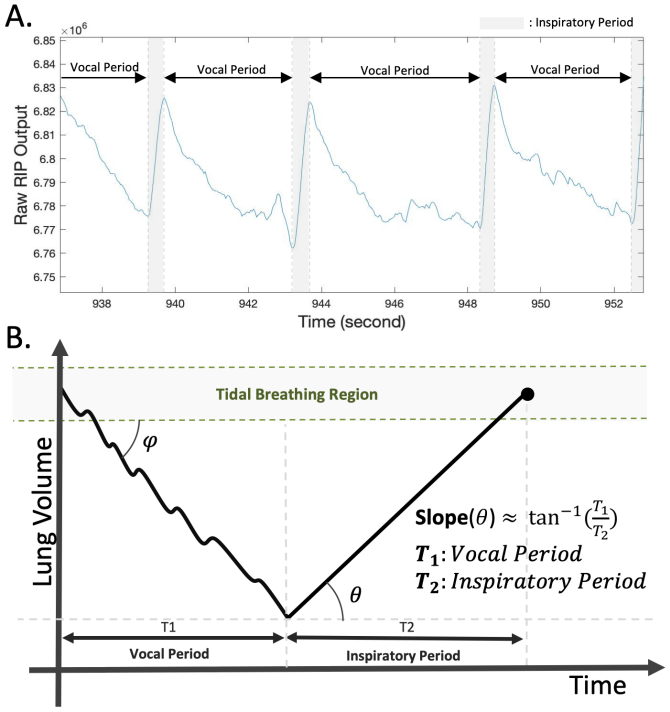

Fig. 2. (A) Actual Respiratory Impedance Plethysmography (RIP) of a participant from the research study is shown here. This plot shows a segment of RIP data during the spontaneous speech task. Slower expiration during vocal period is followed by rapid inspiration. (B) The vocalization to inhalation ratio was approximated by computing the inverse tangent of T_1 to T_2 ratio. One assumption employed in this approximation was that the expiratory rate during speech indicated by the angle φ was constant.

pressure and slowing of the expiratory rate. Due to this high burden in vocalization, patients with obstructive pulmonary disease may present alterations in speaking and swallowing patterns [30]. Specifically, the need for frequent pauses and inspiration imposes limits on the phonation time. Based on this reasoning, we computed the average phonation time from the voice segments and utilized it as one of the features.

7) Inspiration Sound Energy: Previous studies utilizing the inspiratory sound have been done mainly with the auscultation sounds, and their analysis primarily focused on the lower frequency bands (<1,000 Hz) because high frequency components in auscultation sounds are attenuated by the lung and chest cavities, which act as a low pass filter [26]. On the other hand, the inspiration sound collected from the mouth still contains high frequency components. These high frequency components from the inspiratory sounds are generated by the turbulent flows in the airways [26], [31]. In this study, we looked at the frequency band ranging from 7,800 Hz to 8,000 Hz. In this band, the maximum frequency component from the healthy inspiratory sound and pathological inspiratory sound showed meaningful differences, with a larger maximum frequency component in pathological inspiratory sounds.

D. Model Building and Evaluation

We developed two kinds of models from the collected data sets: classification model and regression model. The classification model was designed to detect participants with pathological conditions, and the regression model was built to predict the participant's FEV1-to-FVC ratio, which is one of the standard measures for pulmonary function.

1) Classification Model: The classification model was built with 5 kinds of classifiers (Table I) that are frequently employed in machine learning context. The model development is done using the data set from the research study, and the performance is evaluated in two ways: 10-fold cross-validation for 10 iterations with randomization, and leave-one-participant-out(LOPO). The model developed from the research study data set is evaluated on an independent in-clinic dataset for validating generalizability. The two datasets are described in more detail in Section V.

2) Regression Model: The regression model was built to predict the FEV_1/FVC ratio using the features extracted from the speech data. FEV_1/FVC ratio is considered one of the standard methods utilized in clinics for the assessment of an individual's lung function. According to GOLD guidelines, an FEV_1/FVC less than 0.7 is consistent with an obstructive ventilatory defect [6]. We evaluate the performance of the regression model in terms of mean absolute error (MAE).

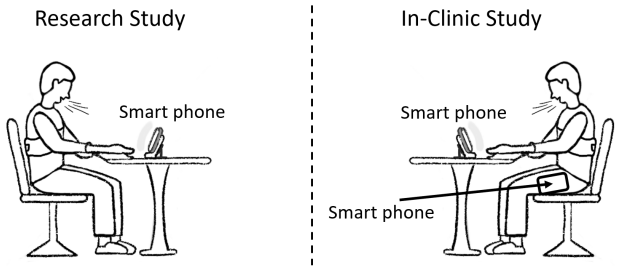

Fig. 3. Study Setup

V. Study Design

Two large-scale studies, a research study and an in-clinic study, were conducted to collect speech data from pulmonary patients and healthy individuals using smartphones. The purpose of the research study was to develop features for a pulmonary disease detection model under acoustically quiet conditions in a research setting. The second study was conducted in collaboration with a hospital to verify the model with a separate data set collected under more realistic, noisy conditions in a clinical visit. Both studies were approved by the Institutional Review Board and a written consent form was obtained from each participant. Both studies shared a similar experimental setup and protocols. Detailed descriptions of each study are provided below.

A. Research Study

1) Participant Recruitment: A total of 131 participants were recruited. Of the 131 (67 males and 64 females) participants, 91 participants had pulmonary conditions and 40

participants were healthy. Among the participants with pulmonary conditions, 69 participants had a history of asthma and 9 participants had a history of COPD. 13 participants reported a history of both asthma and COPD.

2) Measures: The general setup of the research study is described in Figure 3. The audio data were collected from a Samsung Note 8 smartphone. Along with the audio data, the participant's lung function was also measured using a GoSpiro spirometer and respiratory impedance plethysmograph (RIP) from the Zephyr BioHarness 3.0. The spirometer measured the participant's forced expiratory volume in one second (FEV_1) and forced vital capacity (FVC). The RIP band continuously tracked the chest wall excursions throughout the entire study session.

3) Procedures: There was one participant per study session. Upon the participant's arrival, a brief introduction about the study was given and a written consent form was signed by the participant and a researcher. Next, the participant was provided with a Samsung Note 8 smartphone and asked to keep the smartphone either in their hands or on a table within arm's reach while the smart phone was continuously recording audio data. The first task was a pulmonary function test (PFT) using a GoSpiro spirometer. The participant was asked to pinch their nose with the nose clip and inhale deeply followed by maximum exhalation. This process was repeated three times and the best effort was used as the final pulmonary function measurement. The second task was spontaneous speech where the participant was asked to continuously talk for 3-5 minutes on any topics of their choice. After the spontaneous speech task, a scripted speech task began. For this task, the participant was asked to read aloud the 'Rainbow Passage', which is frequently employed in speech analysis due to its phonetic richness. After the scripted speech, the data collection was terminated.

B. In-Clinic Study

1) Participant Recruitment: Initially, a total of 70 participants (60 pulmonary patients and 10 healthy) were recruited from the visitors to the pulmonary clinic who consented to participate. Among the 60 participants with pulmonary conditions, 25 participants had asthma, 10 participants had chronic cough, and 25 participants had COPD. Subsequently, the chronic cough patients were excluded from the classification problem for disease state since the diagnosis for these patients is indeterminate. However, they were included for the regression problem attempting to estimate lung function.

2) Measures: Similar to the research study, the participant's lung function (FEV_1, FVC) was assessed using a clinical grade spirometer (Pneumotrac Portable Screening Spirometer) during the in-clinic study. As in the research study, the smartphone was either held by the participant or placed on the table for recording their speech. One difference between the research study and the in-clinic study was the addition of a Samsung Note 8 smart phone in the participant's pocket (Figure 3). The audio data collected from the smartphone inside the pocket was used to examine the feature robustness for a different acoustic profile that can be expected in a real-world scenario (Section VII-D).

3) Procedures: One subject participated in the study per each session. Prior to the beginning of the study, a brief overview about the study was given and a written consent form was signed by the participant and a researcher. The first task was spontaneous speech. In this task, a smart phone was placed anywhere within arm's reach. The participant was requested to speak continuously for 1 minute on a topic of their choice. The following task was the scripted speech session where the participant was given a passage to read. After 1 minute of reading, the session was stopped. The duration of the speech tasks was reduced compared to the research study in order to be respectful of logistics and patient care in the clinical setting. Subsequently, the participant's pulmonary function was measured using a spirometer. In this step, a certified respiratory technician helped the subject perform spirometry. Three trials were attempted by the participant and the best effort was automatically selected by the spirometer. The FEV_1 and FVC values were recorded.

VI. Results

A. Classification

1) Research Study: Of the 131 participants recruited in the study, 4 healthy participants' data were excluded from the analysis since their FEV_1/FVC ratio was less than 0.7, and one patient's data was not included in the analysis because the speech session was not recorded correctly due to device/app malfunction. In total, data from 126 participants (90 patients, 36 healthy) have been used for the analysis. We built a binary classifier with pathological class (label=1) and healthy class (label=0). The participants who have been diagnosed with obstructive pulmonary diseases such as asthma and COPD were assigned the pathological class. The classifier was examined using two approaches: 10-fold cross-validation and Leave-One-Participant-Out (LOPO).

In 10-fold cross-validation, the process was repeated for 10 iterations with data shuffling in each iteration. We explored several classifiers that are frequently employed in many machine learning applications (Table I). While the best result was obtained using a random forest classifier, gradient boosting classifier and neural network also produced competitive results.

In LOPO, the data from a single participant was left out for testing and the rest of the participants' data were used for training. After iterating through each participant's data for testing, the average result was obtained. With random forest classifier, we obtained an accuracy of 78.6%, F1 of 85.1% (84.6% precision and 85.6% recall), and specificity of 61.1%.

2) In-Clinic Study: An in-clinic study was conducted with a similar protocol, but at a different place and time. Of the initial 70 participants recruited in the clinical study, 9 patients with chronic cough were excluded from classification as mentioned before. In addition, the data from three healthy participants failed to be recorded correctly, and could not be used in

TABLE I
SUMMARY OF 10-FOLD CROSSVALIDATION ON DATA SET FROM RESEARCH STUDY

Classifier	Accuracy	F1	Specificity	AUC
Logistic Regression	0.574	0.670	0.500	0.533
SVM	0.653	0.736	0.677	0.699
Random Forest	0.752	0.829	0.525	0.766
Gradient Boosting	0.725	0.802	0.544	0.725
Neural Network	0.739	0.811	0.530	0.716

TABLE II
SUMMARY OF REGRESSION ANALYSIS IN TERMS OF MAE BETWEEN PREDICTED AND TARGET VALUES

Target	Data Set	Reading	Speech	Reading + Speech
FEV1/FVC	In-Clinic	12.8%	12.5%	12.5%
	Research	8.9%	8.6%	8.9%
	In-Clinic + Research	10.7%	9.7%	9.8%
FEV1%	In-Clinic	20.6%	21.4%	21.2%

the analysis. Lastly, one additional participant's data were excluded because there was a high level of vocal sound in the background from another individual. Even though the ultimate goal is to develop a robust algorithm that can work in all environments even with background speech, we first wanted to validate the features without this challenge. In total, 13 participants' data were not incorporated into the analysis for classification. The classifier trained from the research study was applied to the data from the remaining 57 participants (50 pathological and 7 healthy) in the clinical study. Class imbalance in the in-clinic study was extreme, and thus, the in-clinic study data were decided to be used for cross-study evaluation. In the cross-study evaluation, a random forest classifier was trained on the data set from the research study and tested on the data set from the in-clinic study. With the random forest classifier, we obtained an accuracy of 73.7%, F1 of 84.5% (precision of 87.2% and recall of 82.0%).

B. Regression

1) Research Study: A neural network based regressor was built to predict the FEV_1/FVC ratio, using the same set of features used in the classification problem. The neural network was comprised of two dense layers with 8 units, each followed by dropout layers. From LOPO analysis, a mean absolute error of 8.6 % was obtained (Table II).

2) In-Clinic Study: The same regression analysis approach was applied to the in-clinic data set. The nine participants, who were excluded from the classification analysis due to the possibly non-obstructive nature of the pulmonary symptoms, were included in the regression analysis. Seven participants were not included in the analysis due to missing FEV_1/FVC ratio values for various reasons, including an inability to perform the spirometry correctly. In total, 59 participants were included in the regression analysis. A neural network based regressor was built using the same set of features used in the classification problem. From LOPO analysis, a mean absolute error of 12.5% was obtained (Table II).

VII. DISCUSSION

A. Classification

A binary classifier for distinguishing pathological state (class=1) from healthy state (class=0) was built and evaluated using 1) 10-fold cross validation, 2) LOPO, and 3) cross-study validation. To clarify, the training was only performed on the data set from the research study, and the data set from the in-clinic study was only used for testing and never for training due to the small number of samples and severe class imbalance problem.

1) Cross Validation: Cross validation provides a simple and effective way to evaluate models, and thus, is frequently employed in many machine learning applications today [32]. In our analysis, we employed 10-fold cross validation because it is known to work well on real world data sets [33]. We examined the performance of five different machine learning models as summarized in Table I. While similar performance was obtained from the random forest, gradient boosting, and neural network classifier, the best overall performance was obtained from the random forest classifier with the accuracy of 75.2%, F1 score of 82.9% and AUC of 76.7%.

In addition to the 10-fold cross validation, we also evaluated the model using LOPO cross validation because LOPO analysis is virtually unbiased albeit with high variance [34]. With the random forest classifier, we noticed approximately 3% improvement in accuracy and F1 score, and approximately 9% improvement in specificity (See Section VI-A1). This improvement in classifier performance is likely due to the increased number of samples in the training set.

2) Cross Study Validation: To better evaluate the performance of the model and its generalizability, we trained a classifier using the data set from the research study (n=126) and applied the classifier to the in-clinic data set (n=57). Compared with the results from the LOPO cross validation, the accuracy decreased by 5% and F1 score decreased only by 0.6%. Despite the reduction in accuracy and F1 score, the comparable results between the LOPO cross validation from the research study and the cross-study evaluation are encouraging in terms of the generalizability of the algorithm.

B. Regression

FEV_1/FVC ratio and $FEV_1\%$ are two established metrics for the assessment of obstructive pulmonary diseases. The GOLD (Global Initiative for Chronic Obstructive Lung Disease) guidelines suggest FEV_1/FVC ratio less than 0.7 as the criterion for diagnosing COPD, and subsequently divide COPD severity into four stages based on the percent predicted FEV_1 (a.k.a $FEV_1\%$) [35]. Here, we discuss the regression results for both FEV_1/FVC ratio and $FEV_1\%$.

1) Predicting FEV_1/FVC ratio: Table 2 summarizes the regression error in MAE from LOPO evaluation. Across all trials, a neural network was trained with the 7 features described in Section IV-C to predict FEV_1/FVC ratio. The regression analysis was performed on the nine different combinations of

data sets from the research study and the in-clinic study (Table II). In each study, the data sets from the spontaneous speech session ("Speech" column) and the scripted speech session ("Reading" column) were used to predict FEV_1/FVC ratio (Table II). Then, the data from the two sessions were combined ("Reading + Speech" column) for regression. The data sets were also combined across different studies as indicated by "In-Clinic + Research" in Table II.

From the different combinations of the data sets, we noticed minimal differences in regression error among the reading, the speech and the combined (reading + speech) activities. However, there were noticeable differences in regression error when the two studies were combined; in-clinic study has significantly higher MAE compared with research study. The overall regression error of the combined data sets was higher compared with the regression error from the research study only. We suspect that this decrease in regression performance was due to the greater level of noise present in the in-clinic data set (See Section VII-C), which may have introduced noise to the feature set.

2) Predicting $FEV_1\%$: $FEV_1\%$ was available only from the in-clinic study, and thus, the regression performance was compared using the data set from the in-clinic study. As with FEV_1/FVC ratio described above, there was no significant differences in MAE among reading, speech, and combined data sets. However, the overall MAE of $FEV_1\%$ regression was significantly greater compared with the MAE of the regression for the FEV_1/FVC ratio (Table II). The large difference in MAE between the research study and the in-clinic study may have been due to the large variance of $FEV_1\%$; the standard deviation of FEV_1/FVC ratio in the research study was 0.12 while that of $FEV_1\%$ from the in-clinic study was 0.26.

C. Generalizability

We investigated the generalizability using cross-study evaluation where we trained a model on a data set from the research study and tested the model on the clinical data set. Compared with the 10-fold crossvalidation within the research study data set, there was a significant drop in cross-study evaluation (Table I and Section VI-A2). This difference is most likely caused by the presence of significant noise in the clinical data set. While the research study was conducted in a quiet environment with lesser background noise (-28.8 dB), the in-clinic study took place in a room where there was a constant fan noise in the background as well as other miscellaneous background activity such as door opening/closing and background speech (-10.6 dB). The higher level of the background noise level in the clinical study made it more challenging to identify speech segments and extract reliable features from the data set. Specifically, we noticed that the shimmer and jitter features were noisier in the clinical data sets due to the poor SNR in the audio data. While F1 measure of 76.4% with 82.9% precision and 71.0% recall is encouraging, we believe more robust VAD and feature generation will lead to improved generalizability.

D. Feature Robustness

Speech provides rich information not only about the vocal folds but also about an individual's breathing pattern and potentially airway conditions. However, the analysis of the speech also depends on various factors such as the background noise, distance, and any obstacles between the source and the sensor as such factors may interfere with the feature extraction algorithm. To assess the robustness of the features, we examined the audio files collected from a smartphone in the pocket. Participants from the in-clinic study were asked to put an additional smartphone in their pocket throughout the entire study session, and spontaneous speeches of the participants were recorded simultaneously from the two phones - one on the table within arm's reach and the other in the pocket. We extracted the same set of features described in Section IV-C and applied the classifier trained on the data set from the research study (i.e. cross-study classification analysis).

In this analysis, 55 participants' data were analyzed because the smartphones in the pocket from the two additional participants failed to record audio. As with the previous analysis on the in-clinic data, the random forest classifier trained on the data set from the research study was applied to detect participants with pulmonary diseases. The performance of the classification on the data set collected from the pocket was comparable to the previous result. The accuracy of classification was 70.9%, and the F1 measure was 82.2% (88.1% precision and 77.1% recall). Compared with the classification result from Section VI-A2, accuracy dropped by 2.8% and F1 measure dropped by 2.3%. Based on this comparable accuracy and F1, the proposed algorithm seems robust to the placement of the phone.

E. Feature Selection

Overfitting is one common problem in machine learning applications, which impacts the generalizability of the algorithm. In order to avoid overfitting and improve the generalizability, we selected only a subset from the original set of features listed in Table III. We used recursive feature elimination with cross validation (RFECV) to select the final set of features (Figure 4) [36]. As the number of features increased from 1 to 4, the cross-validation accuracy improved and gradually decreased afterwards albeit with large variability as highlighted by the shaded area which shows the standard deviation. Even though the best accuracy was obtained with 4 features, we decided to use 7 features because an overly small number of features increases the risk of underfitting and reduces robustness to noise. In the selection of 7 features, we used recursive feature elimination (RFE), a feature selection method where weak features are subsequently removed until only specified number of features remain in the feature set [36]. From RFE, we selected the 7 most important features highlighted in Table III. The feature importance of the selected features was computed from the random forest classifier as shown in Figure 5. The vocalization to inhalation ratio, average phonation time, and relative jitter were the top three most important features.

TABLE III
SUMMARY OF INITIAL FEATURES.

Feature #	Feature Description
1	**Pause Frequency (number of pauses / minute)**
2	Shimmer (apq1)
3	Shimmer (apq3)
4	Shimmer (apq5)
5	Shimmer (apq11)
6	**Absolute Jitter**
7	**Relative Jitter**
8	Harmonic-to-Noise Ratio
9	Sum of FFT of inspiratory sound in frequency greater than 2,000 Hz
10	Sum of FFT of inspiratory sound in frequency from 800 Hz to 1,400 Hz
11	**Maximum of FFT of inspiratory sound in frequency from 7,800 Hz to 8,500 Hz**
12	**Mean Inspiratory Slope**
13	Median Inspiratory Slope
14	**Mean of Phonation Period to Inspiratory Period Ratio**
15	**Average Phonation Time**

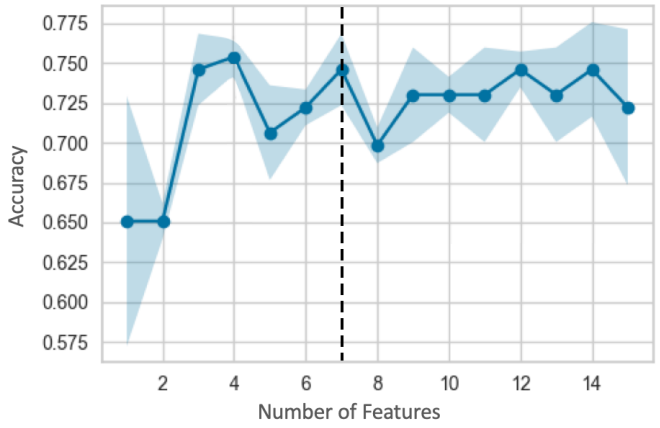

Fig. 4. Recursive feature elimination with cross-validation. The shaded region indicates one standard deviation above and below the mean accuracy.

F. Class Imbalance

In the research study, a total of 131 participants were recruited and 126 participant's data were used in developing and evaluating the model. Among the 126 participant data, however, there was a significant data imbalance problem (36 healthy class and 90 pathological class). Class imbalance causes the classifier to be biased towards the majority class, and thereby, impedes development of an accurate classifier model [37]. There are many well-known approaches for addressing the class imbalance problem, and we chose to address this issue using an oversampling approach.

We used Synthetic Minority Over-sampling Technique (SMOTE) to amplify the samples from the minority class [38]. SMOTE generates synthetic data based on the distribution of the minority class such that the newly generated synthetic data belongs to the distribution of the minority class. Using SMOTE, the number of minority class was matched with the number of majority class during training. SMOTE was not applied to testing. With SMOTE oversampling method, we noticed significant improvement in specificity from the cross validation.

G. Comparison Between Spontaneous and Scripted Speech

Both spontaneous speech and reading involve continuous vocalization. However, spontaneous speech and reading have

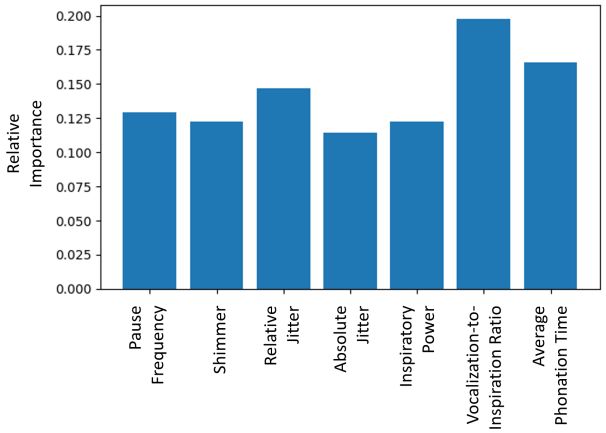

Fig. 5. Feature Importance

different patterns of breath events and different distribution of inspiratory durations [39], [40]. Specifically, there are very short silence periods right before and after the breathing sound in reading [40] whereas spontaneous speech only entails a single silence period [39]. Furthermore, reading activity allows the participant to foresee the sentences to be vocalized, thereby allowing better management of breathing and differed distribution of breathing sounds [39].

As part of the study protocol, the participants were given a set of passages to continuously read out loud for reading task. We selected the passages, like the "Rainbow Passage", that are typically employed by speech pathologists for their phonetic richness. To compare how the performance of our algorithm differs based on type of speech activity, we used the reading audio data collected from the research study. To reiterate, the analysis was performed on the audio data from 126 participants (90 patients and 36 healthy). Random forest classifier was evaluated via 10 repetitions of 10-fold cross-validation, with data shuffling between each repetition. To balance the classes, SMOTE oversampling method was used. The overall accuracy was 62.7% and the F1-score was 73.1% (73.6% precision and 74.6% recall).

VIII. CONCLUSION

We proposed two algorithms that could potentially be applied towards passive assessment of pulmonary function: one for differentiating between healthy individual and those with possible obstructive pulmonary disease, and the other for using speech features to estimate FEV_1/FVC ratio, a clinically accepted pulmonary function metric. From two large data sets collected from mobile devices in controlled settings, the proposed algorithms have been evaluated. Our analysis shows promising results for obstructive pulmonary disease detection and pulmonary function assessment. This work presents a meaningful milestone towards the passive detection of pulmonary disease and the passive assessment of pulmonary functions from spontaneous speech collected from a smartphone.

REFERENCES

[1] J. B. Soriano, A. A. Abajobir, K. H. Abate, S. F. Abera, A. Agrawal, M. B. Ahmed, A. N. Aichour, I. Aichour, M. T. E. Aichour, K. Alam *et al.*, "Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet Respiratory Medicine*, vol. 5, no. 9, pp. 691–706, 2017.

[2] World Health Organization, "Global status report on noncommunicable diseases 2014: attaining the nine global noncommunicable diseases targets; a shared responsability," 2014.

[3] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.

[4] T. Nurmagambetov, R. Kuwahara, and P. Garbe, "The economic burden of asthma in the united states, 2008–2013," *Annals of the American Thoracic Society*, vol. 15, no. 3, pp. 348–356, 2018.

[5] E. S. Ford, L. B. Murphy, O. Khavjou, W. H. Giles, J. B. Holt, and J. B. Croft, "Total and state-specific medical and absenteeism costs of copd among adults aged 18 years in the united states for 2010 and projections through 2020," *Chest*, vol. 147, no. 1, pp. 31–45, 2015.

[6] R. A. Pauwels, A. S. Buist, P. M. Calverley, C. R. Jenkins, and S. S. Hurd, "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Nhlbi/who global initiative for chronic obstructive lung disease (gold) workshop summary," *American journal of respiratory and critical care medicine*, vol. 163, no. 5, pp. 1256–1276, 2001.

[7] S. M. Pollart and K. S. Elward, "Overview of changes to asthma guidelines: diagnosis and screening." *American family physician*, vol. 79, no. 9, 2009.

[8] N. G. Csikesz and E. J. Gartman, "New developments in the assessment of copd: early diagnosis is key," *International journal of chronic obstructive pulmonary disease*, vol. 9, p. 277, 2014.

[9] O. C. van Schayck, A. D'Urzo, G. Invernizzi, M. Román, B. Ställberg, and C. Urbina, "Early detection of chronic obstructive pulmonary disease (copd): the role of spirometry as a diagnostic tool in primary care," *Primary Care Respiratory Journal*, vol. 12, no. 3, p. 90, 2003.

[10] J. C. Zambrano, H. T. Carper, G. P. Rakes, J. Patrie, D. D. Murphy, T. A. Platts-Mills, F. G. Hayden, J. M. Gwaltney Jr, T. K. Hatley, A. M. Owens *et al.*, "Experimental rhinovirus challenges in adults with mild asthma: response to infection in relation to ige," *Journal of allergy and clinical immunology*, vol. 111, no. 5, pp. 1008–1016, 2003.

[11] N. W. Johnston, K. Lambert, P. Hussack, M. G. de Verdier, T. Higenbottam, J. Lewis, P. Newbold, M. Jenkins, G. R. Norman, P. V. Coyle *et al.*, "Detection of copd exacerbations and compliance with patient-reported daily symptom diaries using a smartphone-based information system," *Chest*, vol. 144, no. 2, pp. 507–514, 2013.

[12] Y.-F. Y. Chan, P. Wang, L. Rogers, N. Tignor, M. Zweig, S. G. Hershman, N. Genes, E. R. Scott, E. Krock, M. Badgeley *et al.*, "The asthma mobile health study, a large-scale clinical observational study using researchkit," *Nature biotechnology*, vol. 35, no. 4, p. 354, 2017.

[13] E. J. Sakka, P. Aggelidis, and M. Psimarnou, "Mobispiro: A novel spirometer," in *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*. Springer, 2010, pp. 498–501.

[14] J. A. Walters, R. WOOD-BAKER, J. Walls, and D. P. Johns, "Stability of the easyone ultrasonic spirometer for use in general practice," *Respirology*, vol. 11, no. 3, pp. 306–310, 2006.

[15] C. W. Carspecken, C. Arteta, and G. D. Clifford, "Telespiro: A low-cost mobile spirometer for resource-limited settings," in *2013 IEEE Point-of-Care Healthcare Technologies (PHT)*. IEEE, 2013, pp. 144–147.

[16] S. Gupta, P. Chang, N. Anyigbo, and A. Sabharwal, "mobilespiro: accurate mobile spirometry for self-management of asthma," in *Proceedings of the First ACM Workshop on Mobile Systems, Applications, and Services for Healthcare*. ACM, 2011, p. 1.

[17] E. C. Larson, M. Goel, G. Boriello, S. Heltshe, M. Rosenfeld, and S. N. Patel, "Spirosmart: using a microphone to measure lung function on a mobile phone," in *Proceedings of the 2012 ACM Conference on ubiquitous computing*. ACM, 2012, pp. 280–289.

[19] M. O. Patadia, L. L. Murrill, and J. Corey, "Asthma: symptoms and presentation." *Otolaryngologic Clinics of North America*, vol. 47, no. 1, pp. 23–32, 2014.

[18] M. Goel, E. Saba, M. Stiber, E. Whitmire, J. Fromm, E. C. Larson, G. Borriello, and S. N. Patel, "Spirocall: Measuring lung function over a phone call," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5675–5685.

[20] K. Anderson, Y. Qiu, A. R. Whittaker, and M. Lucas, "Breath sounds, asthma, and the mobile phone," *The Lancet*, vol. 358, no. 9290, pp. 1343–1344, 2001.

[21] S. M. Shaharum, K. Sundaraj, S. Aniza, R. Palaniappan, and K. Helmy, "Classification of asthma severity levels by wheeze sound analysis," in *2016 IEEE Conference on Systems, Process and Control (ICSPC)*. IEEE, 2016, pp. 172–176.

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

[23] M. M. R. E. N. J. K. Viswam nathan, Korosh Vatanparvar, "Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices," in *2019 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2019.

[24] E. Klemmer and F. Snyder, "Measurement of time spent communicating," *Journal of Communication*, vol. 22, no. 2, pp. 142–158, 1972.

[25] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[26] P. Forgacs, "Breath sounds." *Thorax*, vol. 33, no. 6, p. 681, 1978.

[27] M. M. Hassan, M. T. Hussein, A. M. Emam, U. M. Rashad, I. Rezk *et al.*, "Is insufficient pulmonary air support the cause of dysphonia in chronic obstructive pulmonary disease?" *Auris Nasus Larynx*, vol. 45, no. 4, pp. 807–814, 2018.

[28] E. E. Mohamed *et al.*, "Voice changes in patients with chronic obstructive pulmonary disease," *Egyptian Journal of Chest Diseases and Tuberculosis*, vol. 63, no. 3, pp. 561–567, 2014.

[29] B. Wiechern, K. A. Liberty, P. Pattemore, and E. Lin, "Effects of asthma on breathing during reading aloud," *Speech, Language and Hearing*, vol. 21, no. 1, pp. 30–40, 2018.

[30] J. D. Hoit, R. W. Lansing, K. Dean, M. Yarkosky, and A. Lederle, "Nature and evaluation of dyspnea in speaking and swallowing," in *Seminars in speech and language*, vol. 32, no. 01. © Thieme Medical Publishers, 2011, pp. 005–020.

[31] C. M. Rembold and P. M. Suratt, "An upper airway resonator model of high-frequency inspiratory sounds in children with sleep-disordered breathing," *Journal of Applied Physiology*, vol. 98, no. 5, pp. 1855–1861, 2005.

[32] S. Arlot, A. Celisse *et al.*, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[33] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection."

[34] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.

[35] C. F. Vogelmeier, G. J. Criner, F. J. Martinez, A. Anzueto, P. J. Barnes, J. Bourbeau, B. R. Celli, R. Chen, M. Decramer, L. M. Fabbri *et al.*, "Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. gold executive summary," *American journal of respiratory and critical care medicine*, vol. 195, no. 5, pp. 557–582, 2017.

[36] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[37] S. M. A. Elrahman and A. Abraham, "A review of class imbalance problem," *Journal of Network and Innovative Computing*, vol. 1, no. 2013, pp. 332–340, 2013.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[39] T. Fukuda, O. Ichikawa, and M. Nishimura, "Detecting breathing sounds in realistic japanese telephone conversations and its application to automatic speech recognition," *Speech Communication*, vol. 98, pp. 95–103, 2018.

[40] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 838–850, 2007.