2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

SpeechSpiro: Lung Function Assessment from Speech Pattern as an Alternative to Spirometry for Mobile Health Tracking

Korosh Vatanparvar, Viswam Nathan, Ebrahim Nemati, Md Mahbubur Rahman,
Daniel McCaffrey, Jilong Kuang and Jun (Alex) Gao, *IEEE/ACM*

Abstract— **Respiratory illnesses are common in the United States and globally; people deal with these illnesses in various forms, such as asthma, chronic obstructive pulmonary diseases, or infectious respiratory diseases (e.g., coronavirus). The lung function of subjects affected by these illnesses degrades due to infection or inflammation in their respiratory airways. Typically, lung function is assessed using in-clinic medical equipment, and quite recently, via portable spirometry devices. Research has shown that the obstruction and restriction in the respiratory airways affect individuals' voice characteristics. Hence, audio features could play a role in predicting the lung function and severity of the obstruction. In this paper, we go beyond well-known voice audio features and create a hybrid deep learning model using CNN-LSTM to discover spatiotemporal patterns in speech and predict the lung function parameters with accuracy comparable to conventional devices. We validate the performance and generalizability of our method using the data collected from 201 subjects enrolled in two studies internally and in collaboration with a pulmonary hospital. SpeechSpiro measures lung function parameters (e.g., forced vital capacity) with a mean normalized RMSE of 12% and R^2 score of up to 76% using 60-second phone audio recordings of individuals reading a passage.**

Clinical relevance — **Speech-based spirometry has the potential to eliminate the need for an additional device to carry out the lung function assessment outside clinical settings; hence, it can enable continuous and mobile track of the individual's condition, healthy or with a respiratory illness, using a smartphone.**

I. Introduction

Respiratory illnesses or disorders are common in the United States and worldwide, and they come in various forms. Statistics show that 65 million people have moderate to severe chronic obstructive pulmonary disease (COPD) where about 3 million die each year, making COPD the third leading cause of death worldwide [1]. About 334 million people deal with asthma, which is the most common chronic disease in childhood, affecting 14% of children globally [2]. Respiratory tract infections caused by influenza kill between 250,000 and 500,000 people annually. By May 2021, more than a year after the spread of a new strain of coronavirus in 2019, 150 million people were affected by COVID-19 and caused the death of more than 3 million people globally [3]. The United States alone spends billions of dollars on detection, treatment, and management of pulmonary diseases in direct and in-direct healthcare costs [4], [5]. Remote and mobile respiratory health monitoring can potentially provide a patient-centered solution with a high level of care and reduce healthcare costs, mainly decreasing the duration and frequency of subsequent hospitalizations [6], [7].

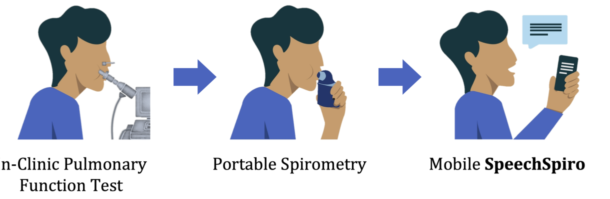

Fig. 1. Transition of the lung function assessment from in-clinic settings to portable devices, mobile spirometry, and towards using natural speech.

Respiratory condition varies day-to-day depending on daily activities, air quality, and environmental factors. Regular assessments are the key to the management of respiratory illness and controlling the spread of infection. Chronic patients may need to take medications daily, and continuous monitoring of relevant symptoms can help clinicians establish if these medications are sufficiently effective [8], [9].

Lung function parameters such as forced vital capacity (FVC) or forced expiratory volume in one second (FEV1) are clinical indices analyzed for respiratory assessment. Changes in these parameters over a couple of days beyond day-to-day variations are an indication of acute exacerbation and worsening condition in the subjects dealing with asthma, COPD, or acute respiratory infections [10]. The decline in lung function would result in a diminished quality of life and cause trouble for subjects with their daily life activities. These individuals may require immediate medical attention and change in their medication to treat the symptoms and avoid worsening conditions. Otherwise, delayed diagnosis and treatment could result in a notable drop in blood oxygen level and even death. Researchers have illustrated that several biomarkers in voice correlate with the respiratory conditions and lung function parameters [8], [11], [12].

Spirometry, as a type of pulmonary function test (PFT) (Figure 1), is a common method to assess lung function [13]. It provides objective information used in the diagnosis of lung diseases and monitoring lung health [14]. Spirometry is a physiological test that measures the maximum air volume that an individual can inhale and exhale with maximal effort. Either volume or flow of air is observed as a function of time, which is later analyzed to identify subjects' airways obstruction or restriction severity. Typically, the PFT parameters such as FVC or FEV1 are normalized into percentages based on the subject's age, race, height, and gender.

K. Vatanparvar, V. Nathan, E. Nemati, Md M. Rahman, D. McCaffrey, J. Kuang, and J. Gao are with Digital Health Lab, Samsung Research America, Mountain View, CA, USA (corresponding author e-mail: korosh.v@samsung.com)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

Human voice production is a complex process to which the respiratory system significantly contributes as it generates the airflow that travels between the vocal folds, thus providing the power source of the voice [15]. As a result of an underlying pulmonary condition, the airflow diminishes due to the obstruction or restriction of the airways, which can compromise the subject's voice or speech [16]. The ubiquity of reliable and efficient microphones in many commodity devices such as smartphones and wearables makes speech relatively convenient to monitor continuously. Hence, it would eliminate the need for individuals to perform a physically intensive task (spirometry) or wear an additional device to assess their lung function.

In this paper, we focus on speech-based spirometry (SpeechSpiro) as an alternative to measuring lung function for mobile health tracking. We create a hybrid deep learning model using a combination of convolutional neural networks and long short-term memory networks (CNN-LSTM). The model extracts spatiotemporal voice patterns in the speech that correlate with obstruction or restriction of the respiratory airways. It analyzes the underlying speech patterns to predict the lung function parameters used in assessing pulmonary conditions and the severity of underlying diseases.

In this work, our main contributions towards speech pattern analysis, speech-based spirometry, and lung function assessment are summarized in the following:

- Detailed analysis of voice/speech audio data and its associated features and patterns that correlate with lung function parameters (e.g., FEV1, FVC, FEV1/FVC) and obstruction of respiratory airways (see Section III-A).
- Hybrid deep learning model using CNN-LSTM to identify localized patterns within short audio segments in frequency domain, extract temporal patterns in speech sequence using the learned high-level latent features, and then correlate with the above-mentioned lung function parameters (see Section III-C).
- Methodology that uses a short period of speech audio recording (e.g., 60 seconds) of an individual reading a passage on a smartphone to assess the lung function assessment - SpeechSpiro (see Section III-B).
- Analyze accuracy and generalizability of the SpeechSpiro methodology compared to medical-grade devices and state-of-the-art approaches using data collected in two lab and clinical studies from 201 subjects of healthy and diagnosed with asthma or COPD with a wide range of severity (see Section IV).

II. Related Work

A pulmonary function test is usually advised for all persons who have respiratory complaints or shortness of breath. In any person, particularly cigarette smokers or asthmatics, spirometry will provide a baseline performance value to track the progression of the condition. PFT is considered as one of the first steps to diagnosis, management, and treatment of respiratory diseases. Different approaches to this test have been introduced which we explain in the following:

Clinical spirometry: Spirometry is typically carried out in a clinical lab setting using medical-grade equipment as the gold standard [17]. It involves a procedure where patients should take a maximal inspiration and forcefully expel air for as long and as quickly as possible. Typically, a pulmonologist or general practitioner performs the test to instruct the subject to put maximum effort and achieve reliable results. A successful test produces a flow-volume curve. With the knowledge of the expected appearance of a flow-volume loop in healthy subjects, a pulmonologist can obtain information of underlying conditions from the morphology of the curve in patients with suspected respiratory disease (see Figure 2). Patients with active respiratory infections are not precluded from having spirometry. However, the tests should ideally be deferred until the risk of cross-contamination is negligible; this could be a limiting factor for lab-based spirometry.

Portable spirometry devices have become more prevalent and provide accessible alternatives to the expensive medical-grade devices outside of the clinical settings, with accuracy comparable to those obtained from the lab-based devices [18], [19], [20]. It enables clinicians to remotely monitor and track the progression of patients' respiratory conditions more frequently. On the other hand, peak flow meters are another inexpensive and portable alternatives to measuring maximum expiratory airflow. However, they mainly reflect flow in the large airways, are effort-dependent, and can be unreliable predictors of asthma exacerbations.

Non-contact spirometry: Discomfort, cost, and availability are the barriers associated with spirometry devices. There have been increasing efforts in developing non-contact respiratory monitoring methods and spirometry to overcome these barriers. For example, new approaches measure respiratory-related chest and abdominal movements and track subtle changes of the chest wall using time-of-flight sensors or active stereo depth-sensing systems comprising a near-infrared (NIR) illuminator and a camera [21], [22], [23]. The displacement of the chest and shoulders are analyzed during the inhalation and exhalation tasks to determine the amount of air volume exchange over time and then mapped to the lung function parameters. The non-contact spirometry approaches demonstrate comparable results to the conventional methods.

Audio-based spirometry: Researchers have demonstrated that besides air pressure during the inhalation and exhalation tasks, audio generated from the mouth correlates with the spirometry parameters [24]. For example, an individual can perform the spirometry task directly using the microphone of a smartphone. Then, an algorithm would measure airflow rate by calculating the envelope of the sound in the time domain to predict lung function.

The above-mentioned conventional approaches to spirometry require the performance of the exhausting spirometry task. Meanwhile, lack of clinical supervision and difficulty with the spirometry task may cause a drop in subjects' compliance, thereby getting unreliable results. This variance is in addition to underlying errors of the measurement due to environmental factors during the spirometry session. These factors include body motion artifacts from the image-based spirometry, background noise, and distance variation of the microphone from the mouth in the audio-based spirometry.

Speech-based spirometry: Data and biomarkers collected from more natural and passive approaches have been investigated as an alternative to overcome the above limitations. For example, researchers have shown a correlation between lung function and audio characteristics from cough sounds or subjects continuously saying the 'A' vowel ('aahh') until they run out of breath [25], [8], [26]. In another approach, which is the primary focus of this paper, researchers analyzed the speech and voice audio features to measure lung function. A set of these features include pause time, fundamental frequency, shimmer, jitter, and their derivatives. These voice features are extracted from audio recordings of an individual speaking for a short period. There has been a significant advancement in developing speech-based spirometry to provide adequate prediction accuracy. However, certain factors still limit the performance of these techniques in a real-world implementation. For example, background noise and low audio quality may compromise the validity of the extracted features such as shimmer or jitter. Moreover, the respiratory condition and severity of the obstruction, cognitive load during speaking may unnoticeably affect the pause time.

In this paper, we address the limitations of the conventional and state-of-the-art approaches. SpeechSpiro is a pulmonary function test that utilizes spatiotemporal speech patterns extracted from audio recordings of a subject reading a paragraph. This methodology can be deployed on smartphones or wearables, eliminating the need for additional devices or sensors. Moreover, there is no need to perform the exhausting inspiratory and expiratory maneuvers in conventional spirometry methods.

III. SpeechSpiro Methodology

Studies have demonstrated that voice characteristics are affected by underlying lung conditions, such as obstruction or restriction in the respiratory airways. In section III-A, correlations between speech and lung conditions are explained further in detail. In section III-C, we leverage these observed correlations and create a hybrid deep learning model to learn and extract spatiotemporal patterns within speech recordings, and then utilize them to measure lung function parameters (section III-B). Figure 3 depicts the high-level methodology and the steps to lung function prediction.

A. Voice Audio and Respiratory Systems

Voice production is mainly carried out in the larynx in three stages. Firstly, vocal fold vibrations create a voiced sound where vocal tract resonators (throat, mouth, and nasal passages) would modulate the voiced sound to produce a human recognizable voice. Finally, vocal tract articulators (tongue, palate, and lips) modify the voiced sound to produce recognizable words or speech. Vocal folds vibrate when excited; air pressure and flow from the lungs control the open and close phases by creating a trailing "Bernoulli effect" [27], [28], [29].

In Figure 2, on the left side, you can see the flow-volume curves during inhalation and exhalation phases of spirometry for subjects with different lung conditions. On the right side, we calculated voice loudness and pitch values for speech and pause segments of all subjects' monologue recordings; the values for these two features are plotted in a 2D histogram - brighter pixels showing values with higher data points. The black line corresponds to the median values of these features for subjects with ”mild” obstruction. The red line corresponds to the median values for subjects with ”very severe” obstruction. The two obstruction severities are identified based on the gold standard severity scale and each subject's pulmonary function test results. As you can see in the figure, the flow-volume curves during inhalation (lower charts) and exhalation (upper charts) show similar patterns comparing to loudness during pause and speech segments.

Pitch and loudness of voice are affected by laryngeal changes and respiratory changes. They tend to co-vary by the airflow and subglottal pressure: the influence of the lungs. An increase in airflow “blows” vocal folds wider apart, which stay apart longer during a vibratory cycle – thus increasing the amplitude of the sound pressure wave or loudness. An increase in the frequency of vocal fold vibration, which is affected by the airflow, raises the pitch. For example, it is easier to sing a high note loudly than softly because of the common factors of airflow and subglottal pressure.

As shown in Figure 2, similar to the expiratory maneuver in a spirometry task, during a speech session, airflow rate naturally decreases as the individual exhales and produces voice. This is due to the fact that the lungs will not have had enough air to provide sufficient subglottal pressure, the longer an individual maintains speaking. This would result in change in loudness and pitch of the voice, typically lower loudness and pitch.

In subjects with underlying lung conditions, an obstruction in the airways limits the airflow rate during exhalation when producing voice (speaking) and during inhalation when grasping for air (pauses between speech) (see Figure 2). The lower inconsistent airflow rate within a speaking session affects the loudness and pitch values of the produced voice, and thereby speech pattern. On the other hand, in the case of underlying inflammation in the lungs, the lung capacity decreases, identified by a low-value FVC during spirometry. The volume of air a subject can hold and their lung capacity impacts the duration they can maintain voice production (speaking). In other words, individuals with small lung capacity may have to talk in shorter sessions and take pauses more frequently in between.

Individuals with a more severe respiratory condition with lower PFT parameters will have more difficulty maintaining their airflow while speaking; hence, the pitch and loudness of their voice will be inconsistent and affected more drastically. This influence and change in the voice audio characteristics can be observed over the period of a speech and pause segment as the individual is freely speaking with frequent pauses in between.

The stat-of-the-art methods heavily rely on average voice

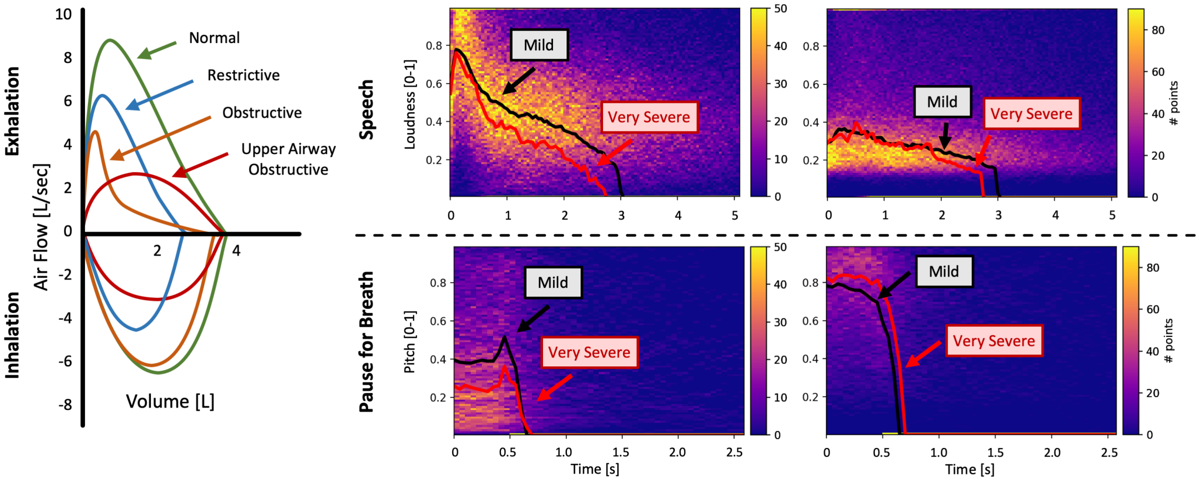

Fig. 2. Airflow-volume curves generated from spirometry session during inhalation and exhalation (on the left side) are compared with pattern of loudness and pitch values during speech and pause segments of the audio recordings (on the right side) for subjects with different pulmonary condition.

audio characteristics of the whole speech session. Instead, in this paper, we leverage the observed changes and patterns of these voice audio characteristics and correlate them with lung function parameters.

B. Speech Pattern Discovery and Modeling

Changes and patterns of voice loudness and pitch values in an audio recording have shown to be correlated with lung obstruction or restriction, and thereby the lung function parameters (see Section III-A). Mel-frequency cepstral coefficients (MFCC) of audio are frequent audio features used in speech-based applications such as speech recognition; they represent the signal power in each band of the frequency domain. The MFCC values can be measured and extracted from audio to extract audio patterns in the frequency and time domain as an alternative to voice loudness and pitch. We consider one of the PFT parameters (e.g., FEV1, FVC, FEV1/FVC) as the prediction target output. Hence, we can formulate the problem as sequential modeling and regression problem that maps the sequence or time-series data of MFCC features to each PFT parameter.

Recurrent neural networks (RNN) with long short-term memory (LSTM) architecture have emerged as an effective and scalable model for sequential data. They are effective in capturing long-term temporal dependencies in a sequence. The model relies on two sources of information to predict future events. One source is derived from a set of recently observed data; the other one is based on a hidden state space defined by the LSTM that aims to abstract past or context information. LSTM variant of RNN has been adapted in different applications such as prediction of rain, precipitation, speech emotion, stock market price movement, and image labels. For the problem of speech-based lung function prediction, the LSTM-based method can be used to learn and extract the temporal dependencies and patterns in the MFCC features. For example, an individual who has limited lung capacity will have difficulty maintaining speech which results in shorter speech activity and consequently having to take longer and more frequent pauses. In other words, an individual who can take a deeper breath during pause time will consequently maintain a longer speech, with more stable voice loudness and pitch. We believe the temporal dependencies between voice characteristics of speech and pause activities through MFCC features represent the subject's underlying lung condition and LSTM can help capture those.

Although LSTM is effective to capture temporal patterns, the redundancy in the fully connected layers may not help capture spatial dependencies. On the other hand, convolutional neural networks (CNN) have been seen as an efficient model to capture spatial and temporal dependencies for classification, localization, and segmentation of one-dimension or multi-dimensional data. CNN architecture is designed such that lower layers fine and detailed features and the higher layers extract more class-specific information. With enough training, CNNs can learn filters and extract characteristics of the data without hand-engineering features. CNN models have been adapted in different applications such as classification and recognition of images, sound classes, natural language processing, and atrial fibrillation (Afib) detection. In our problem, we leverage a 2D CNN to learn and extract spatial and temporal features in the time-series data. For example, as we noticed in section III-A, individuals with severe obstruction would have different patterns of voice loudness and pitch values (or MFCC features) compared to the individuals with normal lungs. Convolutional layers in CNN can learn and extract these patterns that correlate with individuals' underlying lung conditions.

To take advantage of both models of LSTM and CNN, we create a hybrid model that contains LSTM layers following the CNN layers (see Figure 3). The convolutional layers can capture localized spatial and temporal patterns in [sub-sequences of] the time-series data and map them to higher-level localized features for the LSTM layers to identify high-level temporal dependencies in the overall sequence. CNN-LSTM has been recently seen by the research community to be very effective and efficient to learn and model spatiotemporal patterns in sequences.

We use speech audio samples collected in our two research studies to train the model. We have provided details of the data collection procedure, data format, and how we split the dataset for training and validation in section IV-A. Each audio recording of a speech session can vary in duration; hence, the audio is sliced into sequences of 60 seconds with a hop length of 10 seconds. Each sequence of 60s audio is processed to extract a 2D array of time-series MFCC features containing. The features are sampled every 50ms over audio frames with a sliding window size of 100ms.

Each of these feature arrays is input to the CNN-LSTM model for training. To capture the localized spatiotemporal features by the convolutional filters, each sequence of 60s audio and its associated feature arrays are further split into subsequences ($\mathcal{X}_t$) with a duration of 6 seconds and hop length of 3 seconds (50% overlap). Hence, each LSTM cell - hidden state $\mathcal{H}_t$ and cell output $\mathcal{C}_t$ - processes each subsequence to extract localized features and then identify long short-term dependencies over the overall 60s sequence. We have used kernel size of $(5, 3)$ for 2D convolution filters with strides of one. 'ReLu' has been selected as the activation function for the CNN layers and LSTM cells. Since we are interested in one scalar target output, the LSTM layer only outputs one value, the output of the last cell. Finally, a dense layer with a 'tanh' activation function follows the last LSTM layer. The model is trained using *Adam* optimization method and max pooling and drop out layers are utilized properly to avoid overfitting. Model hyperparameters are tuned such that it does not overfit the data. The training process is early terminated after certain epochs and when the difference between training and validation loss values becomes significant.

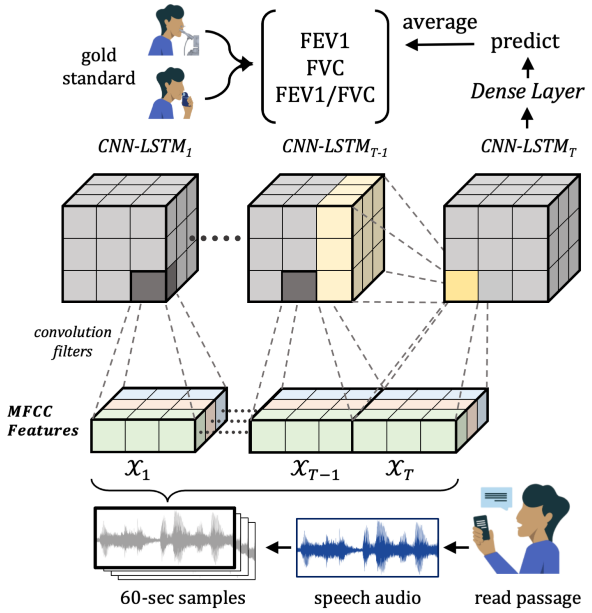

Fig. 3. Diagram of the steps in SpeechSpiro methodology to extract and utilize spatiotemporal speech patterns to predict lung function parameters.

C. Speech-Based Lung Function Assessment

In the SpeechSpiro methodology, a subject will read a passage while a smartphone is recording the audio (see Figure 3). The speech audio is then processed to extract the MFCC features. Similar to the training stage, the features are sampled every 50ms of audio frames with a sliding window size of 100ms. We have noticed that slicing the audio and its features into slices of 60s (hop length of 10s) and taking the average of the predicted outputs would provide better accuracy and less variation in the result. The 2D feature arrays should be sliced into subsequences with a duration of 6 seconds and hop length of 3 seconds (50% overlap) to meet the input dimension of the convolutional filters, similar to the training stage. Finally, the model is applied to the sequence of subsequences to evaluate the lung function parameters. It needs to be noted that there is one individual CNN-LSTM model trained for each lung function parameter.

IV. Experiments

In this section, we describe the experimental setup, dataset, implementation details, baseline models, and then analyze the results and demonstrate the accuracy of our speech pattern-based lung function assessment.

A. Experimental Setup

Datasets. To train and test the models and SpeechSpiro methodology, we used the audio collected in two research studies conducted in lab and clinical settings. Details of the two studies and the data collection are as follows:

1) Lab Study: A total of 131 subjects (67 males and 64 females) were recruited for this study - 40 were healthy individuals and 91 were diagnosed with at least one pulmonary condition, which included 69 subjects with a history of asthma, 9 with COPD, and 13 reported having a history of both asthma and COPD; the cohort of subjects was decided based on self-reporting of their medical history.

In one session, we asked the subjects to read a predefined text - ”Rainbow” passage - out loud for at least 3 minutes. The ”Rainbow” passage is frequently used in speech analyses due to its phonetic richness. The lung function was assessed by a spirometry session performed using a GoSpiro portable spirometry device under the supervision of a research assistant, to ensure maximum effort was put into the test.

2) Clinical Study: A total of 70 subjects were recruited for this study in partnership with a pulmonary hospital - 10 were healthy individuals and 60 were diagnosed with at least one pulmonary condition which included 25 subjects with asthma, 25 with COPD, and 10 with chronic cough; the cohort of subjects was decided based on their medical records. Similar to the lab study, in one session, we asked the participants to read the ”Rainbow” passage out loud for 1 minute. Their lung function was assessed by a spirometry session performed in a PFT lab under the supervision of a pulmonologist, to ensure valid results were achieved.

In both studies, a Samsung Note 8 smartphone located on a table 4ft from the subject was continuously recording audio with a 44.1KHz sampling rate. The audio was later

TABLE I
COMPARING PERFORMANCE OF DIFFERENT SPEECH-BASED LUNG FUNCTION ASSESSMENT METHODS ON TRAINING AND TEST SUBJECT DATA FOR EACH OF THE PFT PARAMETERS.

RMSE [NRMSE: scaled values 0-1] - R2 Score

		Baseline - SVR	CNN	CNN-LSTM
Training Dataset	*FVC*	0.77 [0.16] - 0.23	**0.57 [0.11] - 0.43**	**0.52 [0.10] - 0.76**
	FEV1	0.79 [0.19] - 0.20	**0.79 [0.19] - 0.24**	**0.47 [0.11] - 0.77**
	FEV1/FVC	0.11 [0.18] - 0.11	**0.10 [0.16] - 0.35**	**0.07 [0.10] - 0.72**
Test Subjects Dataset	*FVC*	1.16 [0.24] - 0.11	**1.12 [0.22] - 0.15**	**1.12 [0.22] - 0.22**
	FEV1	1.14 [0.27] - 0.09	**0.99 [0.24] - 0.13**	**1.06 [0.25] - 0.25**
	FEV1/FVC	0.16 [0.25] - 0.03	**0.16 [0.25] - 0.06**	**0.12 [0.20] - 0.22**

segmented and annotated to extract the speech session and label the start and end of each speech and pause segments.

Baseline Methods. We compared multiple machine learning models and analyzed their performance. Here, we show the results for our model comparing with a simpler CNN model and with the state-of-the-art speech-based approach - Support Vector Regression (SVR). In the ”Baseline SVR” approach, we train the model using our dataset on the average values of voice shimmer, jitter, pause time, and MFCC features [8].

Metrics. We calculate the performance in terms of the root mean squared error (RMSE) and determination coefficient R^2. We also provide the normalized RMSE (NRMSE) values for the target outputs scaled to [0-1] for ease of comparison.

Evaluation Strategy. A subset of subjects (10%) has been set aside for testing where we exclude them from the training process; we ensured that the subjects are uniformly split based on their lung condition. Furthermore, we split the data of the training subjects into two sets of training and validation (80%-20%) for unbiased evaluation and tuning of the model to avoid overfitting. Each case study in the paper was repeated three times with shuffled data; we report the average performance in the paper.

B. Experimental Results

We trained the predictive models on the training dataset from both studies and provided the results for both the training dataset and test subject dataset. The results for the three methods and PFT parameters are summarized in Table I. For each case study, the top method with the highest R^2 score appears in bold and underlined; and the second method appears in bold only.

The results show that FVC prediction has the lowest error and highest R^2 score among all PFT parameters. We believe speech pattern and the duration an individual can maintain speaking highly correlates with FVC, which is mainly determined by how much air the individual can hold. Meanwhile, FEV1 requires higher effort in exhalation, which is not always the case during a regular speech.

It is clear that utilizing frequency and temporal patterns in MFCC features, which are extracted from speech, improves the accuracy when comparing the performance of CNN and CNN-LSTM methods with the Baseline SVR method. Furthermore, the temporal information extracted from speech sequence by the CNN-LSTM model improves the overall performance, especially the R^2 score when comparing with the CNN method. This improvement is more prominent for FEV1 prediction and in terms of the achieved R^2 score.

The prediction performance of the CNN-LSTM model against the Baseline method is depicted in Figure 4 by plotting the measured values against the predicted values for the whole dataset.

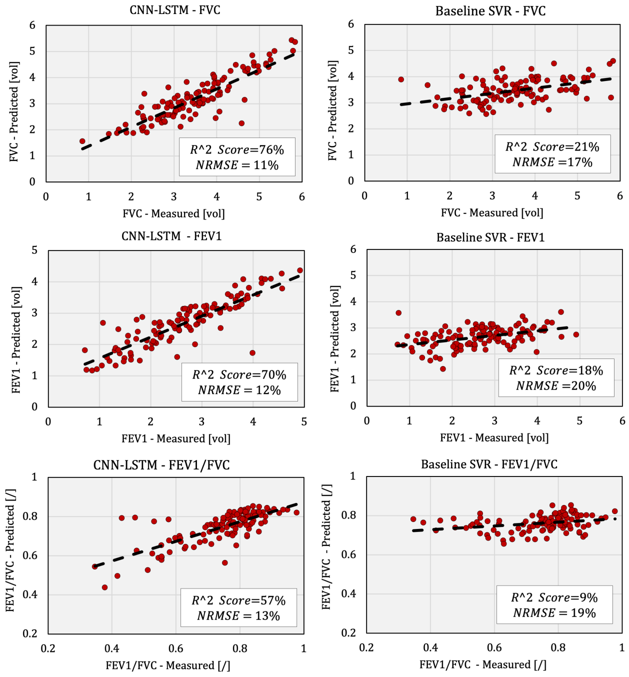

Fig. 4. Regression plots for predicted values versus ground truth values for the whole dataset; regression line is drawn for each parameter that illustrates the difference in R^2 score for both methods.

As shown in the above figures, the speech-based prediction of PFT parameters shows a linear correlation with measured values from the gold-standard spirometry. However, the linear correlation achieved in the CNN-LSTM method shows lower RMSE and much higher coefficient determination with an average NRMSE of 12% and R^2 of up to 76%, compared to the 18% and 21% for the Baseline SVR method.

We further analyzed the accuracy of the SpeechSpiro method (CNN-LSTM) for different cohorts and groups of individuals in terms of their lung condition and gender. As

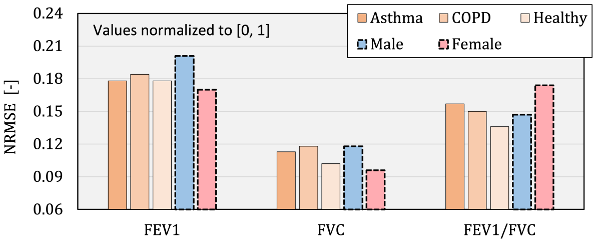

Fig. 5. Performance of the SpeechSpiro methodology (CNN-LSTM) for different categories of individuals in terms of lung condition and gender.

shown in Figure 5, the difference in the mean NRMSE between each group is very negligible as the model is generalizable and capturing the whole range of subjects.

V. Conclusions

In this paper, we explained the limitations of current lung function assessment techniques. We focused on a novel method to extract and leverage speech patterns to predict common PFT parameters. In this method, we developed and utilized a hybrid deep learning model CNN-LSTM to analyze 60 seconds of speech recordings to predict FEV1, FVC, and FEV1/FVC ratio. The method significantly outperformed state-of-the-art techniques and models in terms of mean normalized RMSE of (12%) and determination coefficient (R^2) of up to 76%. We believe, for future work, by utilizing more audio samples, transfer learning, and data augmentation techniques, the performance of the speech-based lung function assessment will further improve. Hence, it has the potential to become a better alternative to conventional spirometry techniques that require an additional device and the performance of an exhausting maneuver.

Acknowledgment

We want to acknowledge Hujun Cui for the phone and watch application development, Leonardo Jimenez Rodriguez, Yoshiya Hirase and Sharath Chandrashekhara for preparing the back-end servers, Dr. Erin Blackstock for data collection supervision, Dr. Christopher Fanta for clinical feedback and the participants for volunteering in the study.

References

[1] W. H. Organization *et al.*, "Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach," in *Global surveillance, prevention and control of chronic respiratory diseases: A comprehensive approach*, 2007, pp. vii–146.

[2] G. A. Network *et al.*, "The global asthma report 2014," *Auckland, New Zealand*, vol. 769, pp. 28–36, 2014.

[3] "Who coronavirus (covid-19) dashboard." [Online]. Available: https://covid19.who.int/

[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] S. M. May and J. T. Li, "Burden of chronic obstructive pulmonary disease: healthcare costs and beyond," in *Allergy and asthma proceedings*, vol. 36, no. 1. OceanSide Publications, 2015, p. 4.

[6] P. P. Walker, P. P. Pompilio *et al.*, "Telemonitoring in chronic obstructive pulmonary disease (chromed). a randomized clinical trial," *American journal of respiratory and critical care medicine*, vol. 198, no. 5, pp. 620–628, 2018.

[7] I. Tomasic, N. Tomasic *et al.*, "Continuous remote monitoring of copd patients—justification and explanation of the requirements and a survey of the available technologies," *Medical & biological engineering & computing*, vol. 56, no. 4, pp. 547–569, 2018.

[8] V. Nathan, M. M. Rahman, K. Vatanparvar *et al.*, "Extraction of voice parameters from continuous running speech for pulmonary disease monitoring," in *International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 859–864.

[9] C. H. Fanta, "Asthma in the elderly," *Journal of Asthma*, vol. 26, no. 2, pp. 87–97, 1989.

[10] M. T. Dransfield, K. M. Kunisaki *et al.*, "Acute exacerbations and lung function loss in smokers with and without chronic obstructive pulmonary disease," *American journal of respiratory and critical care medicine*, vol. 195, no. 3, pp. 324–330, 2017.

[11] V. Nathan, K. Vatanparvar *et al.*, "Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices," in *16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2019, pp. 1–4.

[12] K. San Chun, V. Nathan, K. Vatanparvar *et al.*, "Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone," in *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2020, pp. 1–10.

[13] J. F. Morris, "Spirometry in the evaluation of pulmonary function," *Western Journal of Medicine*, vol. 125, no. 2, p. 110, 1976.

[14] B. L. Graham, I. Steenbruggen *et al.*, "Standardization of spirometry 2019 update. an official american thoracic society and european respiratory society technical statement," *American journal of respiratory and critical care medicine*, vol. 200, no. 8, pp. e70–e88, 2019.

[15] J. Y. Tong and R. T. Sataloff, "Respiratory function and voice: The role for airflow measures," *Journal of Voice*, 2020.

[16] J. D. Hoit, R. W. Lansing *et al.*, "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.

[17] H. Ranu, M. Wilde, and B. Madden, "Pulmonary function tests," *The Ulster medical journal*, vol. 80, no. 2, p. 84, 2011.

[18] W. V. Schneider, B. Bulloch, M. Wilkinson, P. Garcia-Filion, L. Keahey, and M. Hostetler, "Utility of portable spirometry in a pediatric emergency department in children with acute exacerbation of asthma," *Journal of Asthma*, vol. 48, no. 3, pp. 248–252, 2011.

[19] K. Harri, R. S. Tapani, K. Senja, and K. Matti, "Hand-held turbine spirometer: Agreement with the conventional spirometer at baseline and after exercise," *Pediatric allergy and immunology*, vol. 16, no. 3, pp. 254–257, 2005.

[20] K. M. Mortimer, A. Fallot, J. R. Balmes, and I. B. Tager, "Evaluating the use of a portable spirometer in a study of pediatric asthma," *Chest*, vol. 123, no. 6, pp. 1899–1907, 2003.

[21] H. Sakamoto, H. Takamoto *et al.*, "A non-contact spirometer with time-of-flight sensor for assessment of pulmonary function," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 4114–4117.

[22] E. A. Bernal, L. K. Mestha, and E. Shilla, "Non contact monitoring of respiratory function via depth sensing," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2014, pp. 101–104.

[23] C. Liu, Y. Yang *et al.*, "Noncontact spirometry with a webcam," *Journal of biomedical optics*, vol. 22, no. 5, p. 057002, 2017.

[24] E. C. Larson, M. Goel *et al.*, "Spirosmart: using a microphone to measure lung function on a mobile phone," in *Proceedings of the 2012 ACM Conference on ubiquitous computing*, 2012, pp. 280–289.

[25] E. Nemati, M. J. Rahman *et al.*, "Estimation of the lung function using acoustic features of the voluntary cough," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 4491–4497.

[26] A. Rao, E. Huynh, T. J. Royston, A. Kornblith, and S. Roy, "Acoustic methods for pulmonary diagnosis," *IEEE reviews in biomedical engineering*, vol. 12, pp. 221–239, 2018.

[27] S. Björklund and J. Sundberg, "Relationship between subglottal pressure and sound pressure level in untrained voices," *Journal of Voice*, vol. 30, no. 1, pp. 15–20, 2016.

[28] J. Van den Berg, J. Zantema, and P. Doornenbal Jr, "On the air resistance and the bernoulli effect of the human larynx," *The journal of the acoustical society of America*, vol. 29, no. 5, pp. 626–631, 1957.

[29] Z. Zhang, "Effect of vocal fold stiffness on voice production in a three-dimensional body-cover phonation model," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2311–2321, 2017.