PulmoListener: Continuous Acoustic Monitoring of Chronic Obstructive Pulmonary Disease in the Wild

SEJAL BHALLA, University of Toronto, Canada
SALAAR LIAQAT, University of Toronto, Canada
ROBERT WU, University Health Network, Canada
ANDREA S. GERSHON, Sunnybrook Health Sciences Centre, Canada
EYAL DE LARA, University of Toronto, Canada
ALEX MARIAKAKIS, University of Toronto, Canada

Prior work has shown the utility of acoustic analysis in controlled settings for assessing chronic obstructive pulmonary disease (COPD) — one of the most common respiratory diseases that impacts millions of people worldwide. However, such assessments require active user input and may not represent the true characteristics of a patient's voice. We propose PulmoListener, an end-to-end speech processing pipeline that identifies segments of the patient's speech from smartwatch audio collected during daily living and analyzes them to classify COPD symptom severity. To evaluate our approach, we conducted a study with 8 COPD patients over 164 ± 92 days on average. We found that PulmoListener achieved an average sensitivity of 0.79 ± 0.03 and a specificity of 0.83 ± 0.05 per patient when classifying their symptom severity on the same day. PulmoListener can also predict the severity level up to 4 days in advance with an average sensitivity of 0.75 ± 0.02 and a specificity of 0.74 ± 0.07. The results of our study demonstrate the feasibility of leveraging natural speech for monitoring COPD in real-world settings, offering a promising solution for disease management and even diagnosis.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Sejal Bhalla, Salaar Liaqat, Robert Wu, Andrea S. Gershon, Eyal de Lara, and Alex Mariakakis. 2023. PulmoListener: Continuous Acoustic Monitoring of Chronic Obstructive Pulmonary Disease in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 86 (September 2023), 24 pages. https://doi.org/10.1145/3610889

1 INTRODUCTION

The World Health Organization includes chronic respiratory diseases among the four major human chronic diseases[1]. In particular, pulmonary diseases account for an estimated 7.5 million deaths per year, or approximately 14% of annual deaths worldwide [20]. These diseases also impose a major economic burden, with the annual

[1] https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Authors' addresses: Sejal Bhalla, sejal@cs.toronto.edu, University of Toronto, Toronto, Ontario, Canada; Salaar Liaqat, sliaqat@cs.toronto.edu, University of Toronto, Toronto, Ontario, Canada; Robert Wu, Robert.Wu@uhn.ca, University Health Network, Toronto, Ontario, Canada; Andrea S. Gershon, Andrea.Gershon@sunnybrook.ca, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada; Eyal de Lara, delara@cs.toronto.edu, University of Toronto, Toronto, Ontario, Canada; Alex Mariakakis, mariakakis@cs.toronto.edu, University of Toronto, Toronto, Canada.

2474-9567/2023/9-ART86 $15.00
https://doi.org/10.1145/3610889

cost of managing and treating pulmonary diseases increasing steadily over the past decade. Chronic obstructive pulmonary disease (COPD), one of the most common pulmonary diseases [20], is characterized by persistent and progressive airflow limitation [19]. When COPD symptoms significantly worsen within a short amount of time, disruptive exacerbations occur and lead to a faster decline in lung function, poorer quality of life, and increased mortality [46, 54].

Early detection of worsening symptoms is an important step to prevent hospitalization and ensure quicker recovery [62]. Frequent assessment of lung function is therefore necessary for early interventions. The gold standard for assessing lung function is spirometry, which entails blowing air through specialized equipment to measure airflow from the lungs. Spirometry can only be deployed in clinical settings under the active supervision of a trained medical professional, leaving long periods when patients' lung function goes unmonitored. Recent out-of-clinic monitoring techniques have addressed this issue by leveraging the sensing capabilities of pervasive devices such as smartphones and smartwatches to enable frequent lung function assessment. These techniques often require the completion of a strenuous activity like the 6-minute walking test [9] or forced expiration [21, 27] to capture prominent indicators of pulmonary deterioration. These methods impose a burden on the patients, discouraging them from regularly monitoring their condition. Additionally, they present a potential risk of worsening the pulmonary condition for patients with severe illness [11], calling for solutions that rely on less strenuous and more natural alternatives.

Speech can be recorded without significant user effort and is a known indicator of pulmonary health [61]; when air is pushed through the vocal folds to produce speech, it lends useful information about the functional limitations of patients' respiratory pathways. Unlike other respiratory symptoms that occur sporadically, such as coughing or wheezing, speech enables continuous monitoring of lung function. Most existing works using speech rely on active speech recordings collected in controlled laboratory settings [18, 30, 48], but these works are limited in two ways. First, people may unintentionally alter how they speak when asked to do so on command, whether it be by raising their voice or enunciating more clearly than they would during everyday speech. Second, asking people to deliberately record their voice limits the temporal granularity of patient screening. Using speech that is passively recorded throughout the day circumvents these limitations but introduces its own set of unique challenges, including the misinterpretation of background noises and other people's voices as the patient's speech. The background sounds and voices contained in passively recorded audio can also reveal sensitive information, compromising the privacy of the patient and those around them.

In this work, we propose and evaluate a pipeline for passive, continuous speech monitoring during everyday living for assessing COPD symptom severity. Our pipeline, which we call PulmoListener, combines voice activity detection and speaker verification to isolate audio segments that are suitable for assessing the user's pulmonary condition. We then translate the speech segments into sequences of informative features that can be transmitted outside of the patient's device for further analysis while minimizing the leakage of private audio data. To examine the utility of PulmoListener in detecting the worsening of COPD symptoms, we conducted a study with 8 COPD patients who wore smartwatches that passively collected audio over the course of 164 ± 92 days. The patients also self-reported their symptom progression on a daily basis to serve as the labels for our primary analyses. After validating the individual components of the PulmoListener pipeline using auxiliary datasets, we conducted multiple experiments to examine whether PulmoListener can be used to discriminate different levels of COPD symptom severity. We found that PulmoListener is able to classify patients' daily symptom severity with an average sensitivity of 0.79 ± 0.03, specificity of 0.83 ± 0.05, and F1 score of 0.70 ± 0.03 using just 9 days of their own data for model calibration. Moreover, we found that our proposed approach can predict symptom severity up to 4 days in advance with an average sensitivity of 0.75 ± 0.02, specificity of 0.74 ± 0.07, and F1 score of 0.62 ± 0.05.

To summarize, PulmoListener is an end-to-end pipeline that analyzes smartwatch audio recorded during daily living to assess the symptom severity of a COPD patient. We developed PulmoListener with the following objectives:

- **Minimal patient burden:** By processing continuously recorded audio during daily living, PulmoListener circumvents the need for patients to explicitly test themselves on a regular basis. Past work utilising passively collected audio has relied on sporadic respiratory symptoms like coughing [4, 65] and wheezing [3, 8], limiting the temporal granularity of symptom assessment. Instead, PulmoListener uses a voice activity detector and a speaker verification model to isolate short clips of the patient's own speech from ambient noise and non-patient speech. We validated the performance of these components using a real-world dataset and conducted multiple experiments to determine the optimal settings that maximize their accuracy in identifying patient-specific speech.
- **Concise representations of daily speech:** Given the sheer amount of audio that can be collected throughout the day, PulmoListener uses an innovative feature extraction process to summarize a patient's voice in a way that captures temporal variation. This process entails extracting Mel-frequency cepstral coefficients (MFCCs) from non-consecutive speech windows that PulmoListener determines to be most suitable for representing individual hours of the day based on the likelihood that they contain the patient's own speech. The fact that these windows are brief and non-consecutive minimizes the leakage of meaningful speech information, even if a malicious actor were to somehow reconstruct intelligible audio from MFCCs.
- **Classification of current and future COPD symptom severity level:** We collected audio data from 8 COPD patients to evaluate the performance of PulmoListener in classifying daily COPD symptom severity. We provide insights into the importance of personalised models in this context and the minimum amount of data required from new patients for model personalisation. Our results show that PulmoListener can not only detect COPD symptoms on the day that they are reported but also anticipate the worsening of symptoms up to 4 days in advance.

2 RELATED WORK

In this section, we first enumerate various active and passive assessment approaches to pulmonary health assessment. We then draw attention to prior work that has specifically used speech for this purpose.

2.1 Mobile and Wearable Pulmonary Health Assessments

The need for non-invasive and accessible pulmonary assessments has driven the development of mobile and wearable technologies aimed at estimating common respiratory biomarkers. These techniques can be broadly categorized into those that require active user input to assess lung function and those that can passively assess lung function without user intervention.

2.1.1 Active Assessments. One of the gold standards for assessing lung function is spirometry, which entails using a clinical-grade device to measure the capabilities of a person's lungs during forced exhalation. Given the limited convenience and accessibility of spirometry, there are numerous studies that have leveraged smartphone microphones to replicate the same lung function metrics without having to blow through a tube [21, 27, 41, 60]. Regardless, these tests carry the potential risk of exacerbating the lung condition of patients with severe illness due to the vigorous effort required for the maneuver. Similar to forced expiration, voluntary coughs produced on command have been shown to indicate worsened lung function [35, 43, 64]. Saleheen et al. [48] extended this literature by using monosyllabic voice segments (e.g., "ahh" sounds) to estimate lung function. Although these methods have shown great diagnostic potential, they are all highly dependent on user effort and preclude continuous monitoring. Moreover, these studies have been typically been conducted in controlled laboratory settings, limiting their validity in real-world scenarios.

2.1.2 Passive Assessments. Prior work has leveraged multimodal data from ubiquitous devices in an attempt to enable low-effort and continuous assessment of pulmonary health. For example, Tiwari et al. [56] utilized

smartwatch IMU data to monitor heart rate and physical activity over time, achieving an F1 score of 0.41 in classifying binary levels of COPD symptom severity. Other works have utilised audio and IMU data to extract biomarkers derived from respiration rate [40, 42, 44] and breathing phases [25]. These biomarkers have been used to differentiate between healthy individuals and those with respiratory problems, but their direct correlation with pulmonary disease progression remains unclear.

In contrast, acoustic features are more closely related to pulmonary obstruction and could potentially lead to more accurate monitoring. Researchers have built models to detect both coughing [4, 65] and wheezing [3, 8] from passively collected audio, suggesting that these models could be used to generate features for lung function assessment tools. However, not all patients with respiratory diseases present these symptoms, and the frequency of these symptom events may be limited among those who do present with them. The focus of our work, passive speech assessment, circumvents these limitations since most patients speak during daily living.

2.2 Speech Analysis for COPD Monitoring

The complexity of speech production makes it a promising biomarker for health conditions ranging from depression [13] and Parkinson's disease [14] to COVID-19 [22, 67]. Speech is produced by pushing air out of the lungs while shifting vocal folds in a way that modulates airflow; thus, any variation in the functioning of vocal folds or airways is reflected in speech. Prior work has examined the feasibility of analyzing speech recordings to estimate lung function but has done so under a number of assumptions. Researchers have examined the correlation between COPD symptom severity and acoustic features while patients read passages of text [18, 30, 33]. However, read speech can be modulated in a way that does not represent the natural function of a patient's lung function. For example, Nathan et al. [34] note that they chose to have their participants read passages "to avoid any cognitive pauses when the subject tries to think of what to say next [as] these pauses might confound the detection of pauses for breath". Chun et al. [10] explored the use of spontaneous speech for passive lung function assessment, thereby allowing for more natural variations in speech characteristics. However, they conducted their data collection in a quiet indoor environment lacking the confounds that would be encountered by a passive acoustic system operating outside of the clinic, including the presence of background noise and multiple speakers. Closest to our work is that of Sedaghat et al. [50], who account for the aforementioned limitations by developing an end-to-end speech detection and processing pipeline. We improve upon their work by thoroughly examining the outputs of the speech processing components, determining the best constraints for obtaining useful speech, and accounting for temporal speech variations during feature extraction before symptom severity inference.

3 SYSTEM DESIGN

PulmoListener constitutes an end-to-end speech processing pipeline that operates on passively collected audio. The overall pipeline, which is illustrated in Fig. 1, isolates useful speech segments from real-world audio data, converts them into meaningful feature representations, and feeds them into a machine learning model that classifies COPD symptom severity. We describe this pipeline in detail below.

3.1 Preprocessing and Segmentation

To prepare audio recordings for further processing, PulmoListener first removes intervals of silence by sliding a 50-millisecond window and discarding audio that falls below an overall intensity of -20 dB. The non-silent recordings are then split into 2-second windows with 50% overlap. This window size is commonly used in speech analysis literature because it is sufficiently long to capture temporal variations across utterances, leading to sufficient information for steps like speaker verification [23, 53]. To account for the variability that occurs within a 2-second window of speech, each component in the pipeline operates on different frame lengths, but the outputs of those components are aggregated and merged within each window.

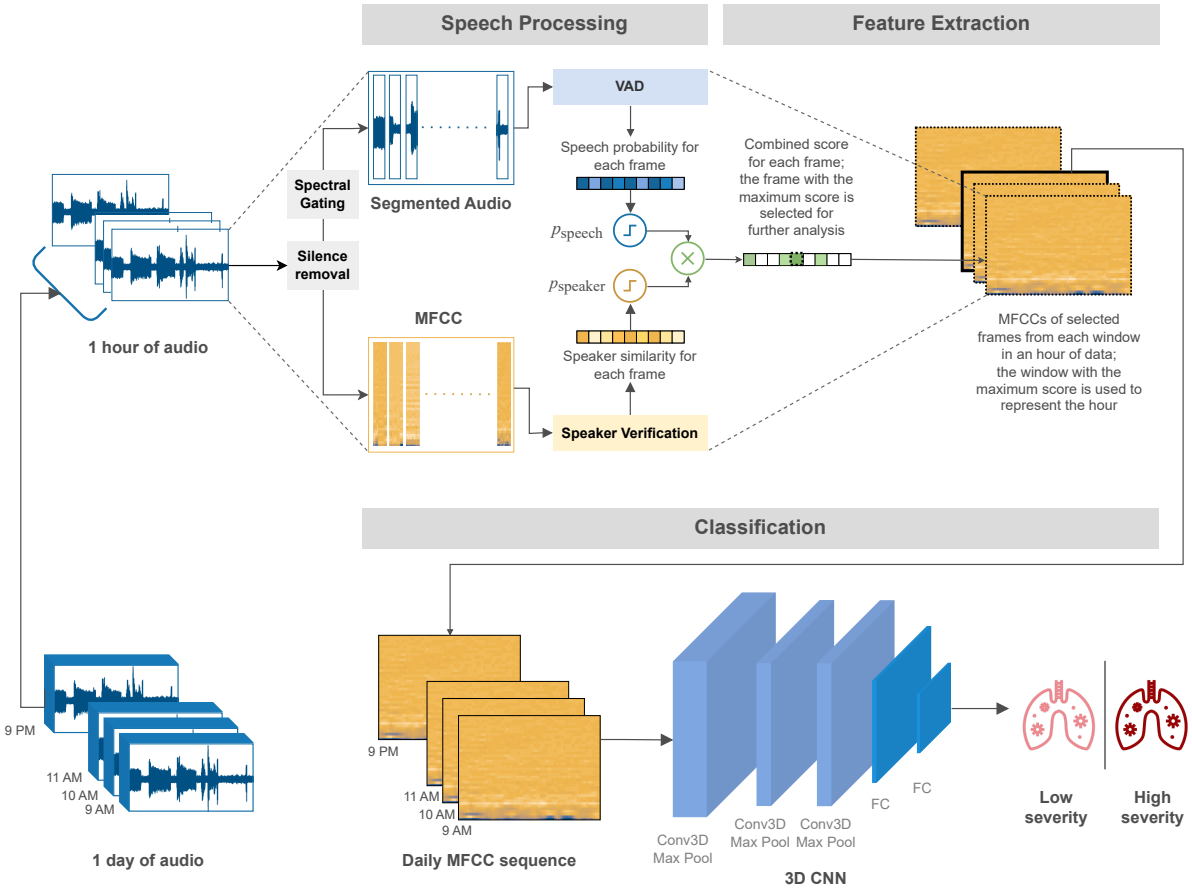


Fig. 1. The system overview of PulmoListener. Audio recorded within a given hour is split, processed, and sorted such that the segment most likely containing the patient's own voice is identified. Features are extracted from those segments across all hours of the day to produce a daily feature vector that can be used for symptom severity classification.

3.2 Voice Activity Detection

For the purposes of this work, any sounds not pertaining to speech are worthless for determining a patient's COPD symptom severity. Thus, it is key to identify voiced sounds in continuous audio recordings. PulmoListener achieves this task using SileroVAD [55], a voice activity detector that operates on audio in the short-time Fourier transform (STFT) domain using a neural network with multi-head attention. SileroVAD is pre-trained on huge corpora that include over 100 languages with various background noise and quality levels. The model is around 1 MB in size, making it suitable for just-in-time prediction on device.

To use this model, PulmoListener suppresses noise in each window using spectral gating [47]. The audio is then split into 20-millisecond frames; SileroVAD is pre-trained with 30-millisecond frames, but shorter windows are equally prominent in speech detection literature [36, 58] and have the advantage of better granularity. We use the model to generate a speech probability score for each frame, which is then binarized using an empirical threshold p_{speech} to separate speech and non-speech segments. The higher the threshold, the more confident

the voice activity detector is that the given frame contains speech; however, this also sacrifices the amount of data left for downstream analyses. We examine this trade-off in Section 6.3.

The binary labels for each frame are aggregated back into their original 2-second windows to identify longer segments of contiguous speech. PulmoListener discards all utterances shorter than 250 ms based on the observation that most meaningful speech utterances are longer than that duration [12]. If the combined duration of all utterances in a window is longer than 1.5 seconds (75% of the window), PulmoListener classifies that window as speech and retains it for further analysis.

3.3 Speaker Verification

Identified speech segments may not only contain the patient's voice, but also the voices of conversational partners or nearby speakers. Excluding such clips is not only important for ensuring that PulmoListener's outputs are representative of the patient, but also as a safeguard for the privacy of others. For this purpose, PulmoListener employs a pre-trained deep residual convolutional neural network (ResCNN) [28] to transform audio into embeddings that can be compared for speaker verification. The model is pre-trained on the LibriSpeech corpus [37], containing over 1k hours of read speech collected from 2.5k speakers. Although the model is trained on English speech, researchers have shown that this model can generalize to other languages using transfer learning [28].

To use this model, PulmoListener extracts Mel-frequency cepstral coefficients (MFCCs) from each 2-second speech window. After being processed by ResCNN, the cosine similarity between the embedding of a test window and a known ground-truth embedding of the patient's voice is used as the speaker similarity score. The ground-truth embedding is obtained by applying the same embedding extraction process to a 2-second window chosen from a patient's clean audio recorded in controlled settings during enrolment. Similar to the voice activity detector, PulmoListener applies an empirical threshold p_{patient} to convert these probabilities into binary labels indicating whether the patient was the speaker in the given clip. We examine the trade-off between confidence and data availability in Section 6.3.

3.4 Feature Extraction

By using speaker verification in conjunction with voice activity detection, the resulting audio segments are assumed to pertain to the patient's own speech and can therefore be used for symptom severity inference. Prior work has shown that changes in a patient's symptom state and their corresponding voice characteristics are gradual, usually hourly or daily [1]. Combined with the fact that COPD symptom severity is self-reported daily at best [63], PulmoListener characterizes symptom severity by processing daily feature sequences.

Within a given hour, PulmoListener discards the audio windows that do not satisfy the thresholds p_{speech} and p_{patient}. The remaining windows are sorted according to the product of their voice activity and speaker similarity scores. The audio windows with the highest overall score for each hour are used for feature extraction and inference. PulmoListener summarizes the patient's speech characteristics within a given day by stacking together the MFCCs of the audio windows selected to represent each of its constituent hours. Since many individuals charge their devices at night, we represent each day using 12 hours' worth of data (9 AM to 9 PM). Whenever fewer than 12 windows are available for generating this sequence, PulmoListener imputes missing windows by selecting unused speech windows from nearby hours.

3.5 Symptom Assessment

The goal of the machine learning model at the end of PulmoListener's pipeline is to distinguish between speech features corresponding to low or high COPD symptom severity. We employed a 3D CNN model for this purpose because CNNs excel at detecting patterns in spatial and temporal sequences [24, 66]. Although recurrent networks

Table 1. The construction of the London COPD Cohort Daily Symptom Questionnaire [1, 52]

Category	Symptom	Score
Major	Increased breathlessness	5
	Change in sputum color	5
	Increased sputum amount	5
Minor	Cold (runny or blocked nose)	1
	Increased wheezing or chest tightness	1
	Sore throat	1
	Worsening cough	1
	Fever	1

such as LSTMs have been used for time-series data, prior work has shown that simple CNNs can often outperform LSTMs at this task [6, 59].

The 3D CNN we use in PulmoListener comprises three convolutional layers with 64, 128, and 128 filters respectively. Each layer is followed by a rectified linear unit (ReLU) activation function, a max pooling layer, and a batch normalisation layer. The output of the last pooling layer is connected to another global average pooling layer and a dense layer with 256 units and ReLU activation function, before the final dense layer with 2 units and softmax activation for classification. The model was trained using the Adam optimizer (learning rate = 0.01, learning rate decay = 0.1) and binary cross-entropy as the loss function.

4 DATASETS

The main dataset we used to evaluate PulmoListener's ability to assess COPD symptom severity consisted of continuous audio and daily symptom responses from COPD patients. The protocol used to collect this data was approved by the Research Ethics Boards at the University Health Network and the University of Toronto under Protocol #41568. Due to privacy restrictions mandated by these entities, we were unable to directly listen to the audio and annotate whether the audio contained speech, let alone the identity of the speaker. Therefore, we used auxiliary datasets to evaluate PulmoListener's speech processing components that required such annotations. All datasets were collected using Samsung Galaxy Watches, thereby eliminating a potential confounding factor in our analyses.

4.1 Main Dataset

This is the primary dataset we used to evaluate PulmoListener and its ability to differentiate between COPD symptom severity levels through everyday speech.

4.1.1 Participants & Protocol. We recruited COPD patients from three hospitals in Toronto, Canada to participate in our study for an average of 164 ± 92 days. Although we recruited 28 patients, we restricted our analyses in this paper to those who had more than 4 weeks of data and exhibited severe symptoms at some point during their enrolment. Applying these inclusion criteria left us with 8 patients (3 females, 5 males), ranging in age from 55–93 years (average = 66.4 ± 11.7 years).

Participants were given a Samsung Galaxy Watch to wear during the day and charge at nighttime. The smartwatch recorded audio data at a sampling rate of 44.1 kHz. To preserve the battery life of the device, a 20% duty cycle was applied such that 2 minutes of continuous audio recording was followed by 8 minutes without

recording. To avoid potentially transmitting sensitive information contained in raw audio, recordings were obfuscated as MFCCs before being uploaded to a centralized server.

Participants reported the severity of their COPD symptoms every morning using the London COPD Cohort Daily Symptom Questionnaire [1, 5] — a clinically validated instrument that has been used in both clinical [51, 52] and technical [29, 50] studies to screen for COPD exacerbations. The questionnaire required patients to reflect on the severity of the eight symptoms listed in Table 1, selecting the ones that were "worse than usual" on the previous day. The questionnaire considers major symptoms as 5 points and minor symptoms as 1 point when tabulating a patient's final symptom score.

4.1.2 Dataset Curation. We were unable to annotate the audio in our dataset due to obfuscation. However, the questionnaire scores required curation so that they could be converted to binary symptom severity labels. The London COPD Cohort Daily Symptom Questionnaire is designed to identify COPD exacerbations, which are serious but rare occurrences in most COPD patients. The questionnaire suggests that scores above 6 for two consecutive days warrant concern; however, applying such a threshold to our dataset led to significant class imbalance during model training. To binarize the scores in our dataset, we adjusted the threshold to 3 and only required patients to exceed this threshold for one day in order to assign a positive label. Although these adjustments deviate from the initial construction of our chosen instrument, their conservative nature also means that PulmoListener would need to be able to identify more subtle manifestations of COPD symptoms.

4.2 Auxiliary Voice Activity Detection Dataset

To evaluate the performance of the voice activity detection model we selected for PulmoListener, we used another dataset from past literature [29] that had smartwatch audio annotated for daily living occurrences (e.g., speech, silence, television sounds). Although this dataset was collected from a different set of participants, the recruitment process, hardware, and protocol were similar to those of the main dataset.

4.2.1 Participants & Protocol. The researchers who collected this dataset recruited 16 COPD patients from three hospitals. There were 12 males and 4 females, and their ages ranged from 52 to 84 (average = 69.3 years). Similar to those in our protocol, the participants in this study were provided with a Samsung Galaxy Watch to collect continuous physiological data and audio. The participants were asked to wear the smartwatch during the day, and the same 20% duty cycling was used to preserve battery life.

4.2.2 Dataset Curation. The researchers on that project hired 10 annotators to localize speech segments within their dataset. To increase the reliability of their labels, each audio segment was assigned to two different annotators. The annotators labeled 38.9 hours of non-silent audio and achieved an average annotator agreement of 77%. In total, roughly 17 hours (43%) of their dataset was determined to contain speech.

4.3 Auxiliary Speaker Verification Dataset

Clean recordings of a person's typical speech are needed in order to register them in a speaker verification model. Since we could not isolate such recordings from our main dataset due to obfuscation, we asked our participants to record their voices in a quiet lab environment at the start of our protocol.

4.3.1 Participants & Protocol. The participants for this dataset were the same COPD patients who completed the protocol for our main dataset. During study enrolment, these individuals recorded voice samples in a quiet room while wearing the Samsung Galaxy Watch that was given to them for continuous data collection. The participants were asked to complete both scripted and unscripted speech tasks. The scripted task entailed reading two short passages that each took an average of 2 minutes to complete, while the unscripted task involved participants speaking continuously on any topic of their choosing for 2 minutes.

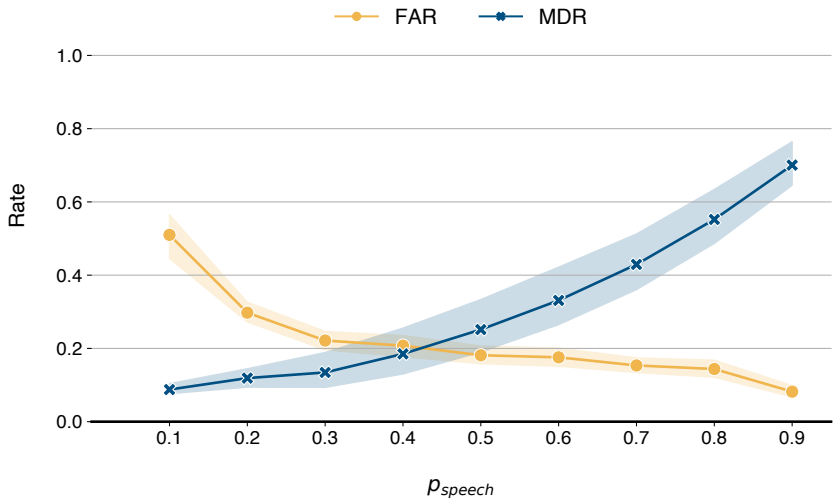


Fig. 2. The performance of SileroVAD [55] on the auxiliary voice activity detection dataset according to false alarm rate (FAR) and missed detection rate (MDR). The confidence intervals indicate the variance in performance across different individuals.

4.3.2 Dataset Curation. Minimal curation was required for this dataset since the protocol was highly controlled and all sound sources were known. We applied a silence threshold to remove segments without speech, after which the remaining segments were labeled with the corresponding participant's identifier.

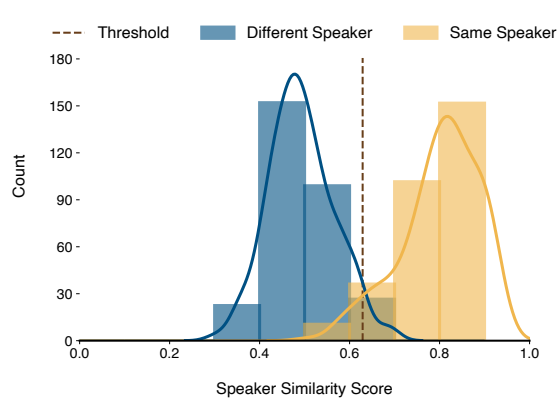
5 VALIDATION OF SPEECH PROCESSING COMPONENTS

In this section, we report the performance of the selected voiced activity detection and speaker verification models on our auxiliary datasets.

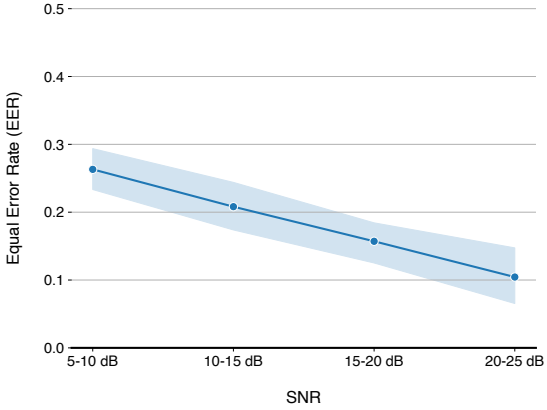
5.1 Voice Activity Detection

To assess the degree to which SileroVAD can identify voiced segments in labeled real-world audio data collected from smartwatches, we evaluated its performance on the auxiliary voice activity detection dataset described in Section 4.2. We used the model to generate a speech probability score for each audio frame in all of the recordings, after which the scores were converted to binary labels according to the threshold p_{speech}. We report SileroVAD's performance using two common metrics used for quantifying the performance of voice activity detection models: false alarm rate (FAR) and missed detection rate (MDR). FAR is defined as the proportion of non-speech incorrectly classified as speech, whereas MDR is defined as the proportion of speech incorrectly classified as non-speech. We report the variance of these metrics across participants to highlight the varied effectiveness of SileroVAD on different individuals.

Fig. 2 illustrates SileroVAD's performance as a function of p_{speech}. As expected, higher settings of p_{speech} led to decreased FAR and increased MDR, and lower settings of p_{speech} led to the opposite trends. At a threshold of $p_{\text{speech}} = 0.5$, SileroVAD was able to identify speech with an average false alarm rate of 0.18 ± 0.03 and an average missed detection rate of 0.21 ± 0.06. Moving the threshold to $p_{\text{speech}} = 0.3$ minimized overall error even further, but we consider $p_{\text{speech}} = 0.5$ to be the more optimal setting in this experiment given the greater importance of FAR over MDR in the context of PulmoListener. Regardless, it does not follow that either of these



(a) The distribution of speaker similarity scores for clean speech recordings from the same speaker and different speakers. The vertical dashed line indicates the threshold that optimally separates the two distributions.



(b) The equal error rate (EER) achieved on synthetically modified speech recordings. The confidence intervals indicate the variance in performance across different individuals.

Fig. 3. The performance of Res-CNN [28] on the auxiliary speaker verification dataset.

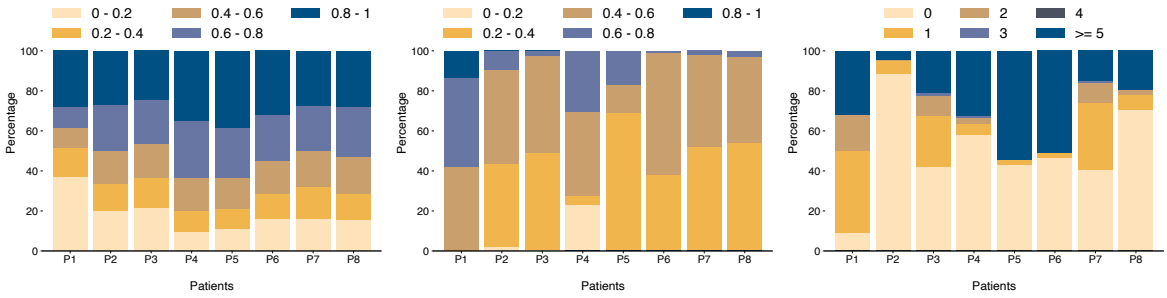
thresholds is best for PulmoListener. As p_{speech} increases, the suitability of the remaining speech segments for speaker verification and symptom severity inference will also increase. However, increasing the threshold also means that PulmoListener may not have enough viable segments to properly account for variations in speech characteristics throughout the day. We examine these considerations in Section 6.3.

5.2 Speaker Verification

To assess the ability of Res-CNN to differentiate the voices of patients in our dataset from other unidentified speakers, we evaluated its performance on the auxiliary speaker verification dataset described in Section 4.3. Each recording in the dataset was compared against 40 randomly selected recordings from the same patient and 40 randomly selected recordings from other patients. We used the model to generate a similarity score for each recording, after which the scores were converted to binary labels according to the threshold p_{patient}. We report Res-CNN's performance in terms of equal error rate (EER), which describes the model's performance when false acceptance rate and false rejection rates are equal.

The similarity score distributions generated during this experiment are shown in Fig. 3a. Although there is overlap between the distributions of positive and negative speech examples, we found that the model could achieve an average EER of 0.09 ± 0.06 across patients when p_{patient} was optimally personalised for each individual. The optimal setting for p_{patient} across all patients was 0.63. With this setting, Res-CNN was still able to achieve a reasonable EER of 0.12. As with the voice activity detector, there are other factors beyond the performance of the speaker verification model that determine the optimal setting of p_{patient} for PulmoListener, and we examine these considerations in Section 6.3.

5.2.1 Robustness to Real-World Noise. To anticipate the susceptibility of our speaker verification model to real-world noise, we evaluated Res-CNN's performance on synthetic noisy data. We added noise to clean speech recordings in the auxiliary speaker verification dataset by mixing them with background sounds sourced from



(a) Speech probability scores (b) Speaker similarity scores (c) Symptom scores

Fig. 4. The distribution of speech probability scores, speaker similarity scores, and symptom scores for each patient in the main dataset.

the Environmental Sound Classification dataset [39]. The latter dataset consists of audio recordings organized into five categories: indoor (e.g., HVAC), outdoor (e.g., birds chirping), urban (e.g., vehicle traffic), animal (e.g., dogs barking), and social (e.g., cafe) background noise. By randomly selecting 5 background sounds from each category, we created 25 new noisy audio clips for every clean audio clip. Although it would have been desirable to generate noisy clips that matched the signal-to-noise ratio (SNR) of the real-world audio in our main dataset, we could not establish such a target without a deterministic noise signal in the real-world audio. However, Pearsons et al. [38] report that people are able to maintain a 5–8 dB SNR when conversing outside their homes, a 9–14 dB SNR when talking inside their homes, and around 7 dB SNR in public places. Thus, we generated noisy audio with varying SNR levels from 5 to 25 dB in intervals of 5 dB (i.e., 5–10 dB, 10–15 dB, etc.).

Fig. 3b illustrates the performance of the model as a function of SNR with respect to our synthetically modified speech recordings. Decreasing the SNR of the audio increased EER, but even at an SNR of 5 dB, the model was able to verify speakers' identities with an EER of 0.26 ± 0.03.

6 VALIDATION OF PULMOLISTENER

In this section, we first describe the characteristics of our main dataset that are relevant for interpreting our experiment results. We then present the experiments we conducted to identify the optimal configuration of PulmoListener. After examining the importance of model personalisation with respect to symptom assessment, we investigate the possibility of using PulmoListener for prediction days ahead rather than on the same day. Note that all descriptive statistics related to data availability assume that the silence removal step in PulmoListener has already been applied to the audio.

6.1 Descriptive Statistics

Speech probability scores and speaker similarity scores are two key features that dictate the assumed suitability of data for symptom severity inference. Although we were unable to annotate whether audio from our main dataset came from patients' own voices, Fig. 4a and Fig. 4b show the distributions of the corresponding scores within each patient. Patients generated an average of 2.07 ± 0.13 hours of non-silent audio each day. The average speech probability score across all patients was found to be 0.56 ± 0.29, while the average speaker similarity score across all patients was 0.49 ± 0.09. The range of both scores was nearly 1.00, indicating clear differences across audio windows according to these metrics. The large variance in speech probability scores is a result of the different sounds that people encounter in their daily lives. In contrast, the speaker similarity scores were generally low due

Table 2. The size and performance of potential model architectures for symptom severity inference in PulmoListener. For the purposes of this experiment, p_{speech} and p_{patient} were both set to 0.5.

Model	Training Parameters	Sensitivity	Specificity	F1 Score
SVM	-	0.55 ± 0.07	0.52 ± 0.05	0.42 ± 0.03
Random Forest	-	0.58 ± 0.04	0.55 ± 0.06	0.49 ± 0.03
LSTM	935k	0.71 ± 0.09	0.69 ± 0.07	0.64 ± 0.07
3D CNN	223k	**0.78 ± 0.06**	**0.82 ± 0.07**	**0.70 ± 0.04**

to two reasons: (1) the reference audio recording for each patient was collected in a controlled environment as opposed to our main dataset which was collected in the wild, and (2) many of the scores correspond to segments that likely did not contain speech.

Patients reported their symptoms over a combined total of 1,310 days, with an average score of 2.7 ± 4.2 across all reports. As illustrated in Fig. 4c, the distribution of these scores within each patient points to a significant imbalance in the dataset. Thus, we binarized the symptom scores into "low" or "high" symptom severity levels based on whether the symptom score exceeded a threshold of 3.

6.2 Model Selection for Symptom Severity Inference

To assess whether continuous speech holds sufficient information to detect elevated symptom severity, we examined a variety of feature extractors and model architectures for symptom severity inference:

- **Machine learning:** As a baseline approach, we featurised the dataset using the COMPARE feature extractor from the openSMILE toolkit [17]. This feature extractor calculates 6.3k acoustic features, including MFCCs, using diverse functionals over low-level descriptor contours [49]. For each feature, we computed seven aggregates over the day: mean, median, maximum, minimum, skewness, kurtosis, and standard deviation. Since this resulted in a large number of features, we selected the top 200 features according to their Gini importance.

(1) **SVM:** C = 10, γ = 0.1, kernel = radial basis function
(2) **Random forest:** Number of trees = 100

- **Deep learning:** We featurised the dataset using PulmoListener's pipeline for MFCC generation. These sequences were then rearranged according to the input size of the network.

(3) **LSTM:** Number of layers = 2, number of units per layer = [64, 128]
(4) **3D CNN (PulmoListener):** Number of convolutional layers = 3, number of units per convolutional layer = [64, 128, 128], number of dense layers = 2, number of units per dense layer = [256, 2]

All of our models were implemented in Python using `scikit-learn`[2] and `Keras`[3] with their default values unless otherwise specified.

Since patients reported their symptoms on a daily basis, each day of data collection yielded a single training sample consisting of an MFCC sequence and the corresponding symptom severity level. This feature extraction scheme resulted in a small dataset with as many samples as the number of days for which patients were enrolled in the study. To obtain enough data for training deep learning models, we augmented our main dataset by generating ten MFCC sequences per day with a relaxed criterion for window selection in each hour. First, the top ten windows $\{w_i^1, w_i^2, \ldots, w_i^{10}\}$ for each hour i were selected based on the product of their speech probability and speaker similarity scores. A window was then selected from each hour without replacement to form a sequence

[2] https://scikit-learn.org/stable/index.html
[3] https://keras.io/

such as $\{w_1^3, w_2^2, \ldots, w_{12}^6\}$. This process was repeated until all the selected windows in an hour had been used to form a sequence, resulting in ten distinct sequences that captured variations in patients' speech characteristics within a day.

To maximize the likelihood of success, we built personalised models for each patient by training on all other patients' data and two weeks of the patient's own data. We ensured that all sequences from the same day were kept in the same split to prevent information leakage. Before model training, the dataset was passed through the voice activity detection and speaker verification models with their respective thresholds set to 0.5.

Table 2 reports the classification performance of all the models along with the number of trainable parameters. We found that the deep learning architectures outperformed the traditional machine learning alternatives, likely due to a combination of model complexity and their ability to model sequential data. The 3D CNN outperformed the rest of the model architectures with an average sensitivity of 0.78 ± 0.06, specificity of 0.79 ± 0.07, and F1 score of 0.70 ± 0.04. Therefore, we proceeded to use the 3D CNN for our subsequent experiments.

6.3 Threshold Selection

Next, we examined the implications of varying p_{speech} and p_{patient} within the broader context of PulmoListener. These thresholds determine the amount and quality of data that is used to train the symptom severity inference model, therefore impacting PulmoListener's overall performance. To conduct this experiment, we followed the same procedure as before to train personalised models for each patient. However, we varied the data used to train these models by trying different settings for the speech processing stages. We used a grid-search approach such that $p_{\text{speech}} \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ and $p_{\text{patient}} \in \{0.0, 0.2, 0.4\}$. PulmoListener's performance when p_{speech} and p_{patient} were set to 0.0 served as a baseline without any speech processing. We did not explore higher values of p_{patient} because there was not enough audio satisfying this threshold to warrant model training. We also did not explore lower values of p_{speech} besides the baseline configuration since running speaker verification on non-speech audio is counterproductive.

Table 3 summarizes the results for each pair of threshold settings. Without any speech processing, PulmoListener had an average sensitivity of 0.59 ± 0.03, specificity of 0.61 ± 0.05, and F1 score of 0.52 ± 0.04. Unsurprisingly, introducing both processing components and increasing their thresholds led to decreased data availability, with changes in p_{speech} typically having greater impact than changes in p_{patient}. There were no monotonic trends with respect to either threshold on its own. For settings of p_{speech} at or below 0.2, varying p_{patient} had little impact on symptom severity inference. This observation reflects the fact that running speaker verification on non-speech audio is counterproductive since such models are typically only trained on speech segments. When p_{speech} was set to 0.4, increasing p_{patient} improved model performance across all metrics. For higher settings of p_{speech}, however, model performance degraded as p_{patient} increased and data became more sparse. Thus, the best model configuration from our experiments was $p_{\text{speech}} = 0.4$ and $p_{\text{patient}} = 0.4$, achieving an average sensitivity of 0.77 ± 0.03, specificity of 0.77 ± 0.02, and F1 score of 0.69 ± 0.04. Using a conservative threshold for p_{speech} allowed PulmoListener to rule out audio windows that clearly contained non-speech sounds. Meanwhile, using a p_{patient} setting that was more aggressive given the distribution of speaker similarity scores in our dataset served its purpose of ensuring that only segments pertaining to the patient's own speech remained. In other words, the speaker verification model likely excluded audio windows containing either background noise or speech utterances from other speakers. For the rest of our experiments, we set $p_{\text{speech}} = 0.4$ and $p_{\text{patient}} = 0.4$.

6.4 Personalisation

Although subject-independent models are ideal for their scalability, they often suffer from poor generalisability owing to idiosyncrasies in people's voices. Therefore, we examined the utility of model personalisation wherein

Table 3. The availability of audio data that satisfied various values of p_{speech} and p_{patient} and the subsequent impact of those threshold selections on the performance of PulmoListener.

Speech Probability Score (p_{speech})	Speaker Similarity Score (p_{patient})	Fraction of Available Audio Windows per Patient	Sensitivity	Specificity	F1 Score
0.0	0.0	100.00 ± 0.00%	0.53 ± 0.03	0.61 ± 0.05	0.52 ± 0.04
	0.2	96.67 ± 4.57%	0.55 ± 0.11	0.57 ± 0.03	0.45 ± 0.06
	0.4	68.24 ± 6.14%	0.56 ± 0.07	0.57 ± 0.04	0.45 ± 0.04
0.2	0.0	81.59 ± 5.61%	0.56 ± 0.04	0.59 ± 0.07	0.47 ± 0.05
	0.2	80.13 ± 4.82%	0.56 ± 0.06	0.58 ± 0.04	0.46 ± 0.04
	0.4	62.36 ± 4.92%	0.58 ± 0.03	0.60 ± 0.06	0.50 ± 0.05
0.4	0.0	68.37 ± 3.44%	0.69 ± 0.02	0.69 ± 0.03	0.61 ± 0.05
	0.2	66.14 ± 2.64%	0.70 ± 0.04	0.73 ± 0.03	0.63 ± 0.03
	0.4	51.59 ± 4.32%	**0.77 ± 0.03**	**0.77 ± 0.02**	**0.69 ± 0.04**
0.6	0.0	46.27 ± 2.03%	0.72 ± 0.03	0.72 ± 0.04	0.62 ± 0.02
	0.2	44.26 ± 3.17%	0.73 ± 0.02	0.72 ± 0.03	0.63 ± 0.01
	0.4	38.52 ± 2.42%	0.75 ± 0.03	0.77 ± 0.02	0.66 ± 0.03
0.8	0.0	30.49 ± 3.27%	0.65 ± 0.03	0.65 ± 0.02	0.58 ± 0.02
	0.2	29.67 ± 2.52%	0.66 ± 0.04	0.67 ± 0.03	0.58 ± 0.02
	0.4	27.64 ± 1.94%	0.62 ± 0.03	0.64 ± 0.02	0.55 ± 0.02

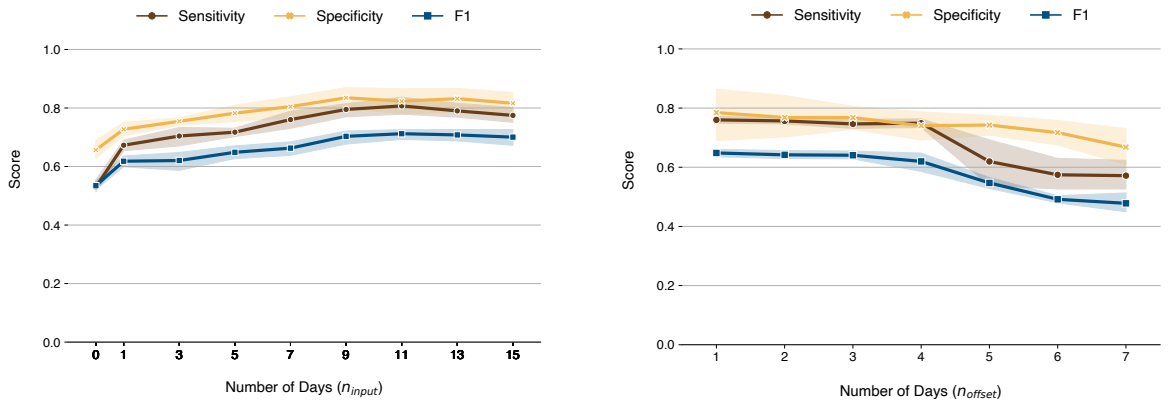
models trained with other patients' data are calibrated for new patients using a few days of their data. In order to determine the amount of labeled data required to achieve reasonable performance with minimal patient burden, we trained the model on the first $n_{\text{input}} \in \{1, 3, 5, \ldots, 15\}$ days of their data. The remaining days of the patient's data comprised the test set used for evaluation.

Fig. 5a shows how the inclusion of subject-specific data improved model performance. Without any personalisation, PulmoListener achieved an average sensitivity of 0.60 ± 0.04, specificity of 0.66 ± 0.05, and F1 score of 0.53 ± 0.02. These metrics increased to 0.79 ± 0.03, 0.83 ± 0.05, and 0.70 ± 0.03 respectively with 9 days of data. Although PulmoListener performed even better with 11 days of subject-specific training data, we consider this increase to be minimal relative to the patient effort required for two extra days of symptom reports. Thus, we use 9 days of audio and corresponding symptom reports to personalize the symptom severity inference model for each new patient.

6.5 Prediction of Daily Symptom Severity in Advance

Prior work has shown that changes in voice characteristics can be observed a few days before the patient starts exhibiting symptoms [32], which could help patients and clinicians anticipate exacerbations and engage in proactive healthcare. To investigate how far in advance PulmoListener can identify potential symptom worsening, we trained models to predict the severity level $n_{\text{offset}} \in \{1, 2, 3, \ldots, 7\}$ days later.

The average performance across all patients for different values of n_{offset} is shown in Fig. 5b. As expected, the predictive capability of PulmoListener decreased as the model was designed to anticipate elevated symptoms further in advance. Nevertheless, PulmoListener's accuracy remained stable when n_{offset} was extended as far



(a) The impact of using more days of patient data to personalise PulmoListener.

(b) The ability of PulmoListener to predict elevated symptom severity days in advance.

Fig. 5. The performance of PulmoListener in detecting and forecasting symptom severity level.

as 4 days in advance, achieving an average sensitivity of 0.75 ± 0.02, specificity of 0.74 ± 0.07, and F1 score of 0.62 ± 0.05.

7 DISCUSSION

In the following subsections, we discuss the implications of our work, acknowledge its limitations, and highlight potential directions for future work.

7.1 Key Findings & Implications

Overcoming several real-world challenges through our end-to-end speech processing pipeline, we demonstrate the feasibility of using audio collected from smartwatches during daily living to characterize COPD symptom severity. Our systematic evaluation of PulmoListener highlights the utility of each of its constituent components. We performed a thorough comparative analysis of the outputs generated by the speech processing components to identify the best thresholds for extracting useful speech, improving the F1 score with respect to symptom severity inference from 0.52 ± 0.04 to 0.69 ± 0.04. Personalising the classifier with a few days of patient data further enhanced PulmoListener's performance, with a notable increase in sensitivity and specificity as compared to patient-independent models. Our optimal configuration, which uses a speech probability threshold p_{speech} = 0.4, a speaker similarity threshold p_{patient} = 0.4, and 9 days of patient data for model personalisation allowed PulmoListener to classify patients' daily symptom severity level with an average sensitivity of 0.79 ± 0.03, specificity of 0.83 ± 0.05, and F1 score of 0.70 ± 0.03. With a marginal reduction in these performance metrics, PulmoListener can also anticipate the deterioration of COPD symptoms up to 4 days in advance.

The results of this work, combined with the convenience of using passively collected audio, have several implications for personal COPD symptom tracking and remote patient monitoring. PulmoListener can be deployed in a way that notifies patients of their deteriorating pulmonary condition, empowering them to proactively avoid triggers and to make lifestyle changes for better disease management. Continuous monitoring also allows healthcare providers to supervise patients' health remotely, reducing the need for frequent in-person visits or daily self-reports. This benefit can be especially helpful for elderly or disabled patients who might have difficulty

traveling to clinics. Overall, these affordances can give patients and their healthcare providers ample time to employ preventive treatment in a way that leads to a quicker recovery and slower disease progression [62].

7.2 Acoustic Noise Artifacts

Passively collected audio is susceptible to noise that can interfere with the suitability of recorded speech for analysis. In some cases, background noise can mask or muddle certain speech sounds; in other instances, background noise can be misconstrued as speech altogether. To address these issues, prior work has suggested combining techniques such as noise suppression, independent component analysis, speech recognition, and speaker verification to isolate and remove noise [16, 26, 45, 57]. Given the overlapping frequency content of speech and ambient sounds, noise suppression and source separation can distort speech in a way that makes it difficult to detect changes in vocal characteristics. We instead relied on voice activity detection and speaker verification to identify speech segments that could be readily classified as speech, and we examined the trade-offs between data quality and availability with respect to the scores output by these components. Still, these techniques may have discarded potentially useful data for analysis, so exploring processing techniques for borderline speech segments may be an opportunity for further exploration.

7.3 Privacy Concerns

Privacy is a major concern for any passive acoustic sensing system. People often mention sensitive information in their conversations, and even background noises can implicate their location or surroundings. We mitigated this risk by leveraging privacy-preserving MFCCs to process a patient's speech for all but one component of PulmoListener. During the MFCC extraction process, the phase information is lost when computing the power spectrum and the Mel scale quantizes the magnitude spectrum into a reduced resolution [2]. The result is a lossy reconstruction that is distorted and typically unintelligible. These factors make it difficult to extract meaningful speech information from MFCC features alone [31]. The only part of our pipeline that did not use MFCCs was the voice activity detector, which instead utilized STFT features since MFCCs are known to be susceptible to noise [7].

Still, we envision that our models for voice activity detection and speaker verification could be deployed on the patient's personal device with minimal modification. This would mean that only daily MFCC sequences would need to be transmitted to a centralized server for symptom severity inference, as illustrated Fig. 6. These sequences are composed of 12 MFCCs extracted from 2-second audio windows sampled from different hours throughout the day. In other words, they only contain 24 seconds of non-consecutive speech data from each patient daily. Transmitting and storing such aggregated MFCC sequences on an external device entails minimal leakage of meaningful speech information, even if a malicious actor were to somehow reconstruct intelligible audio from them.

7.4 Other Indicators of Pulmonary Dysfunction

The findings of our work demonstrate the feasibility of using vocal characteristics to model lung function deterioration. However, the vocalization process is not limited to speech but rather encompasses the production of breathing sounds, wheezing, and coughing as well. We relied strictly on audio pertaining to a patient's speech for assessing their pulmonary symptoms given the relative quantity, duration, and prominence that speech has relative to other body sounds. The faintness of inhalation and the low frequency of wheezing and coughing made them less attractive as target events for PulmoListener. Nevertheless, the success of recent cough detection techniques [65] and breathing phase monitors [25] holds promise for comprehensive analysis of a patient's vocal characteristics.

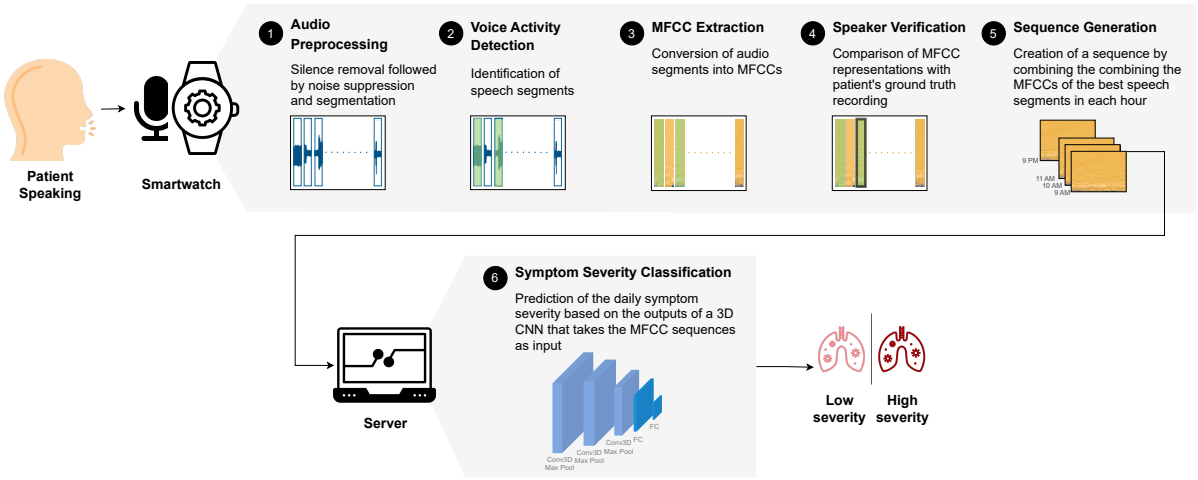


Fig. 6. An illustration of how PulmoListener could be deployed across both a smartwatch and a server to minimize the leakage of sensitive speech audio. All components other than the symptom severity classification model would be implemented on the patient's smartwatch, leaving only aggregated sequences of non-consecutive MFCCs to be transmitted to the server.

Respiratory disorders are also often accompanied by autonomic dysfunction and comorbidities such as myocardial infarction. Thus, vital signs such as heart rate and oxygen saturation have also been widely used to measure complementary aspects of pulmonary condition [56]. Using present-day smartwatches, reliable estimates of most of these vital signs can be easily obtained and processed for analysis. We believe that a multimodal yet ubiquitous approach to analyzing multiple streams of data would result in better pulmonary assessments during daily living, and we plan on investigating the efficacy of such a model in future work.

7.5 Limitations

There are some limitations in our work that affect the ecological validity of our results. First, we were unable to annotate our main dataset given its size and the considerations of participant privacy. We instead relied on auxiliary datasets to validate our voice activity detection and speaker verification components. We assumed that the demonstrated performance of the voice activity detection model would translate to our main dataset since the corresponding auxiliary dataset was collected using similar hardware during daily living. We also assumed that the demonstrated performance of the speaker verification model would translate to our main dataset since the corresponding auxiliary dataset involved the same hardware and target users. Even so, there are undoubtedly differences between the characteristics of these datasets.

Second, our use of the self-reported symptom reports to define the final output of PulmoListener leaves significant room for further exploration. As dictated by the London COPD Cohort Daily Symptom Questionnaire [1, 52], the scores we used to quantify COPD severity accounted for multiple symptoms, some of which may be more directly related to lung function. We also binarized the scores to account for the skewed and unique distribution of symptoms exhibited by our patient participants. Although framing our task as a binary decision of overall symptom severity provides useful insights into the manifestation of COPD, we intend on exploring more fine-grained tasks like symptom score regression or individual symptom classification in future work. These tasks would require the use of oversampling or augmentation techniques to create a balanced dataset.

Third, we did not comprehensively investigate the impact of every possible hyperparameter in PulmoListener's pipeline. For example, we summarized a day's worth of audio data using hour-long intervals in order to generate a manageable feature representation while preserving some degree of variation over time. Our choice for interval length was based on prior literature stating that voice characteristics change gradually over the course of hours [1]. Nevertheless, adjusting this hyperparameter impacts PulmoListener in various ways, including the number of input features and the quality of speech windows that are selected for prediction. We also compared different machine learning and deep learning architectures to select the optimal model for our dataset, but other model architectures may also prove to yield better performance.

Lastly, we evaluated PulmoListener using audio collected by smartwatches due to their convenience, affordability, and proximity to vocalization sources. The adoption of smartwatches has stagnated in recent years [15], so researchers could explore other endpoints for passive acoustic sensing. Smartphones would be the most obvious option, but we decided against this approach since they are often kept in people's pockets or purses. Another option would be to leverage smart speakers, which have the advantage of being kept in environments that are reasonably quiet with only a handful of speakers.

8 CONCLUSION

Although there has been significant interest in speech analysis for assessing pulmonary health, existing approaches have relied on active audio recordings collected in controlled laboratory settings. These approaches can only be leveraged during periods when patients are willing to record themselves and are susceptible to unintentional changes in vocalization when people are asked to speak on command, calling for methods that support passive assessment. Our work represents a significant step to this end, examining the importance of voice activity detection and speaker verification when analysing audio captured during daily living. In its optimal configuration, we found that PulmoListener achieved an average sensitivity of 0.79 ± 0.03, specificity of 0.83 ± 0.05, and F1 score of 0.70 ± 0.03 when classifying patients' symptom severity on the same day. We also found that PulmoListener can predict symptom severity up to 4 days in advance, which could provide benefits for patients who wish to anticipate elevated symptoms in advance. We hope that our work provides a foundation upon which other researchers can explore opportunities for passive audio analysis in the context of health and beyond.

ACKNOWLEDGMENTS

We acknowledge the funding support of Samsung Research America and Natural Sciences and Engineering Research Council of Canada (NSERC) (funding reference number RGPIN-2021-03457 and RGPIN-2017-06618).

REFERENCES

[1] Shawn D Aaron, Gavin C Donaldson, George A Whitmore, John R Hurst, Tim Ramsay, and Jadwiga A Wedzicha. 2012. Time course and pattern of COPD exacerbation onset. Thorax 67, 3 (March 2012), 238–243. https://doi.org/10.1136/thoraxjnl-2011-200768

[2] Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. 2022. Mel Frequency Cepstral Coefficient and its Applications: A Review. IEEE Access 10 (2022), 122136–122158. https://doi.org/10.1109/ACCESS.2022.3223444

[3] Mohsin Y Ahmed, Md Mahbubur Rahman, Viswam Nathan, Ebrahim Nemati, Korosh Vatanparvar, and Jilong Kuang. 2019. mLung: Privacy-Preserving Naturally Windowed Lung Activity Detection for Pulmonary Patients. In 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, Chicago, IL, USA, 1–4. https://doi.org/10.1109/BSN.2019.8771072

[4] Forsad Al Hossain, Andrew A Lover, George A Corey, Nicholas G Reich, and Tauhidur Rahman. 2020. FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, 1 (2020), 1–28.

[5] NR Anthonisen, J Manfreda, CPW Warren, ES Hershfield, GKM Harding, and NA Nelson. 1987. Antibiotic therapy in exacerbations of chronic obstructive pulmonary disease. Annals of internal medicine 106, 2 (1987), 196–204.

[6] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 4 (Dec. 2021), 1–20. https://doi.org/10.1145/3494994

[7] Utpal Bhattacharjee, Swapnanil Gogoi, and Rubi Sharma. 2016. A statistical analysis on the impact of noise on MFCC features for speech recognition. In 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE, Jaipur, India, 1–5. https://doi.org/10.1109/ICRAIE.2016.7939548

[8] Soujanya Chatterjee, Md Mahbubur Rahman, Tousif Ahmed, Nazir Saleheen, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, and Jilong Kuang. 2020. Assessing Severity of Pulmonary Obstruction from Respiration Phase-Based Wheeze-Sensing Using Mobile Sensors. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376444

[9] Qian Cheng, Joshua Juen, Shashi Bellam, Nicholas Fulara, Deanna Close, Jonathan C. Silverstein, and Bruce Schatz. 2017. Predicting Pulmonary Function from Phone Sensors. Telemed J E Health 23, 11 (Nov. 2017), 913–919. https://doi.org/10.1089/tmj.2017.0008

[10] Keum San Chun, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Md Mahbubur Rahman, Erin Blackstock, and Jilong Kuang. 2020. Towards Passive Assessment of Pulmonary Function from Natural Speech Recorded Using a Mobile Phone. In 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE, Austin, TX, USA, 1–10. https://doi.org/10.1109/PerCom45495.2020.9127380

[11] B. G. Cooper. 2011. An update on contraindications for lung function testing. Thorax 66, 8 (Aug. 2011), 714–723. https://doi.org/10.1136/thx.2010.139881

[12] Ruth E. Corps, Birgit Knudsen, and Antje S. Meyer. 2022. Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. Cognition 223 (2022), 105037. https://doi.org/10.1016/j.cognition.2022.105037

[13] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. Speech Communication 71 (July 2015), 10–49. https://doi.org/10.1016/j.specom.2015.03.004

[14] Biswajit Das, Khalid Daoudi, Jiri Klempir, and Jan Rusz. 2019. Towards Disease-specific Speech Markers for Differential Diagnosis in Parkinsonism. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brighton, United Kingdom, 5846–5850. https://doi.org/10.1109/ICASSP.2019.8683887

[15] Deloitte. 2018. Smartwatch adoption rate among consumers in the United States from 2013 to 2018. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-global-mobile-consumer-survey-extended-deck-2018.pdf.

[16] J. Droppo and A. Acero. 2004. Noise robust speech recognition with a switching linear dynamic model. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1. IEEE, Montreal, Que., Canada, I–953–6. https://doi.org/10.1109/ICASSP.2004.1326145

[17] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia. ACM, Firenze Italy, 1459–1462. https://doi.org/10.1145/1873951.1874246

[18] Mireia Farrús, Joan Codina-Filbà, Elisenda Reixach, Erik Andrés, Mireia Sans, Noemí Garcia, and Josep Vilaseca. 2021. Speech-Based Support System to Supervise Chronic Obstructive Pulmonary Disease Patient Status. Applied Sciences 11, 17 (Aug. 2021), 7999. https://doi.org/10.3390/app11177999

[19] Global Initiative for Chronic Obstructive Lung Disease. 2019. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease 2019 report. https://goldcopd.org/wp-content/uploads/2018/11/GOLD-2019-v1.7-FINAL-14Nov2018-WMS.pdf.

[20] Roger I. Glass and Joshua P. Rosenthal. 2018. International Approach to Environmental and Lung Health. A Perspective from the Fogarty International Center. Annals ATS 15, Supplement_2 (April 2018), S109–S113. https://doi.org/10.1513/AnnalsATS.201708-685MG

[21] Mayank Goel, Elliot Saba, Maia Stiber, Eric Whitmire, Josh Fromm, Eric C. Larson, Gaetano Borriello, and Shwetak N. Patel. 2016. SpiroCall: Measuring Lung Function over a Phone Call. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, San Jose California USA, 5675–5685. https://doi.org/10.1145/2858036.2858401

[22] Jing Han, Chloe Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2021. Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Toronto, ON, Canada, 8328–8332. https://doi.org/10.1109/ICASSP39728.2021.9414576

[23] Hee-Soo Heo, Youngki Kwon, Bong-Jin Lee, You Jin Kim, and Jee-weon Jung. 2022. High-resolution embedding extractor for speaker diarisation. https://doi.org/10.48550/ARXIV.2211.04060

[24] Andrey Ignatov. 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Applied Soft Computing 62 (Jan. 2018), 915–922. https://doi.org/10.1016/j.asoc.2017.09.027

[25] Bashima Islam, Md Mahbubur Rahman, Tousif Ahmed, Mohsin Yusuf Ahmed, Md Mehedi Hasan, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Jilong Kuang, and Jun Alex Gao. 2021. BreathTrack: Detecting Regular Breathing Phases from Unannotated Acoustic Data Captured by a Smartphone. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 3 (Sept. 2021), 1–22. https://doi.org/10.1145/3478123

[26] Ju-ho Kim, Jungwoo Heo, Hye-jin Shim, and Ha-Jin Yu. 2022. Extended U-Net for Speaker Verification in Noisy Environments. (2022). https://doi.org/10.48550/ARXIV.2206.13044 Publisher: arXiv Version Number: 1.

[27] Eric C. Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N. Patel. 2012. SpiroSmart: using a microphone to measure lung function on a mobile phone. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, Pittsburgh Pennsylvania, 280–289. https://doi.org/10.1145/2370216.2370261

[28] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. 2017. Deep Speaker: an End-to-End Neural Speaker Embedding System. (2017). https://doi.org/10.48550/ARXIV.1705.02304 Publisher: arXiv Version Number: 1.

[29] Daniyal Liaqat, Salaar Liaqat, Jun Lin Chen, Tina Sedaghat, Moshe Gabel, Frank Rudzicz, and Eyal de Lara. 2021. Coughwatch: Real-World Cough Detection using Smartwatches. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Toronto, ON, Canada, 8333–8337. https://doi.org/10.1109/ICASSP39728.2021.9414881

[30] Julia Merkus, Ferdy Hubers, Catia Cucchiarini, and Helmer Strik. 2020. Digital Eavesdropper -Acoustic Speech Characteristics as Markers of Exacerbations in COPD Patients.

[31] Ben P. Milner, Jonathan Darch, and Ibrahim Almajai. 2009. Reconstructing clean speech from noisy MFCC vectors. In Interspeech.

[32] Mir Mohammed Daanish Ali Khan, Prakhar Pradeep Naval, Rajat Kulshreshtha, Satya Venneti, and Anil Singh. 2021. VOICE-BASED MONITORING OF COPD. Chest 160, 4 (Oct. 2021), A2173–A2174. https://doi.org/10.1016/j.chest.2021.07.1920

[33] Venkata Srikanth Nallanthighal, Aki Harma, and Helmer Strik. 2022. Detection of COPD Exacerbation from Speech: Comparison of Acoustic Features and Deep Learning Based Speech Breathing Models. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, Singapore, 9097–9101. https://doi.org/10.1109/ICASSP43922.2022.9747785

[34] Viswam Nathan, Korosh Vatanparvar, Md Mahbubur Rahman, Ebrahim Nemati, and Jilong Kuang. 2019. Assessment of Chronic Pulmonary Disease Patients Using Biomarkers from Natural Speech Recorded by Mobile Devices. In 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, Chicago, IL, USA, 1–4. https://doi.org/10.1109/BSN.2019.8771043

[35] Ebrahim Nemati, Md. Juber Rahman, Erin Blackstock, Viswam Nathan, Md. Mahbubur Rahman, Korosh Vatanparvar, and Jilong Kuang. 2020. Estimation of the Lung Function Using Acoustic Features of the Voluntary Cough. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, Montreal, QC, Canada, 4491–4497. https://doi.org/10.1109/EMBC44109.2020.9175986

[36] Kuldip K. Paliwal, James G. Lyons, and Kamil K. Wojcicki. 2010. Preference for 20-40 ms window duration in speech analysis. In 2010 4th International Conference on Signal Processing and Communication Systems. IEEE, Gold Coast, Australia, 1–4. https://doi.org/10.1109/ICSPCS.2010.5709770

[37] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[38] K Pearsons, R Bennett, and S Fidell. 1977. Speech levels in various noise environments.

[39] Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM International Conference on Multimedia (Brisbane, Australia) (MM '15). Association for Computing Machinery, New York, NY, USA, 1015–1018. https://doi.org/10.1145/2733373.2806390

[40] Md Mahbubur Rahman, Mohsin Yusuf Ahmed, Tousif Ahmed, Bashima Islam, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2020. BreathEasy: Assessing Respiratory Diseases Using Mobile Multimodal Sensors. In Proceedings of the 2020 International Conference on Multimodal Interaction. ACM, Virtual Event Netherlands, 41–49. https://doi.org/10.1145/3382507.3418852

[41] Md Mahbubur Rahman, Tousif Ahmed, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, Erin Blackstock, and Jilong Kuang. 2020. ExhaleSense: Detecting High Fidelity Forced Exhalations to Estimate Lung Obstruction on Smartphones. In 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE, Austin, TX, USA, 1–10. https://doi.org/10.1109/PerCom45495.2020.9127355

[42] Md. Mahbubur Rahman, Ebrahim Nemati, Viswam Nathan, and Jilong Kuang. 2020. InstantRR: Instantaneous Respiratory Rate Estimation on Context-Aware Mobile Devices. In 13th EAI International Conference on Body Area Networks. Springer International Publishing, 267–281. https://doi.org/10.1007/978-3-030-29897-5_22

[43] Vishwajith Ramesh, Korosh Vatanparvar, Ebrahim Nemati, Viswam Nathan, Md Mahbubur Rahman, and Jilong Kuang. 2020. CoughGAN: Generating Synthetic Coughs that Improve Respiratory Disease Classification. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, Montreal, QC, Canada, 5682–5688. https://doi.org/10.1109/EMBC44109.2020.9175597

[44] Ruth Ravichandran, Elliot Saba, Ke-Yu Chen, Mayank Goel, Sidhant Gupta, and Shwetak N. Patel. 2015. WiBreathe: Estimating respiration rate using wireless signals in natural settings in the home. In 2015 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE, St. Louis, MO, USA, 131–139. https://doi.org/10.1109/PERCOM.2015.7146519

[45] Chandan K A Reddy, Anshuman Ganguly, and Issa Panahi. 2017. ICA based single microphone Blind Speech Separation technique using non-linear estimation of speech. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, New Orleans, LA, 5570–5574. https://doi.org/10.1109/ICASSP.2017.7953222

[46] Kieran J. Rothnie, Hana Müllerová, Liam Smeeth, and Jennifer K. Quint. 2018. Natural History of Chronic Obstructive Pulmonary Disease Exacerbations in a General Practice–based Population with Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 198, 4 (Aug. 2018), 464–471. https://doi.org/10.1164/rccm.201710-2029OC

[47] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS computational biology 16, 10 (2020), e1008228.

[48] Nazir Saleheen, Tousif Ahmed, Md Mahbubur Rahman, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, Erin Blackstock, and Jilong Kuang. 2020. Lung Function Estimation from a Monosyllabic Voice Segment Captured Using Smartphones. In 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, Oldenburg Germany, 1–11. https://doi.org/10.1145/3379503.3403543

[49] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In Interspeech 2016. ISCA, 2001–2005. https://doi.org/10.21437/Interspeech.2016-129

[50] Tina Sedaghat, Salaar Liaqat, Daniyal Liaqat, Robert Wu, Andrea Gershon, Tatiana Son, Tiago H. Falk, Moshe Gabel, Alex Mariakakis, and Eyal de Lara. 2022. Unobtrusive Monitoring of COPD Patients using Speech Collected from Smartwatches in the Wild. In 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). IEEE, Pisa, Italy, 818–823. https://doi.org/10.1109/PerComWorkshops53856.2022.9767283

[51] Terence AR Seemungal, Gavin C Donaldson, Elizabeth A Paul, Janine C Bestall, Donald J Jeffries, and Jadwiga A Wedzicha. 1998. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine 157, 5 (1998), 1418–1422.

[52] Terence A. R. Seemungal, Gavin C. Donaldson, Angshu Bhowmik, Donald J. Jeffries, and Jadwiga A. Wedzicha. 2000. Time Course and Recovery of Exacerbations in Patients with Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine 161, 5 (May 2000), 1608–1613. https://doi.org/10.1164/ajrccm.161.5.9908022

[53] Yanpei Shi, Mingjie Chen, Qiang Huang, and Thomas Hain. 2020. T-vectors: Weakly Supervised Speaker Identification Using Hierarchical Transformer Model. https://doi.org/10.48550/ARXIV.2010.16071

[54] Samy Suissa, Sophie Dell'Aniello, and Pierre Ernst. 2012. Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. Thorax 67, 11 (Nov. 2012), 957–963. https://doi.org/10.1136/thoraxjnl-2011-201518

[55] Silero Team. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

[56] Abhishek Tiwari, Salaar Liaqat, Daniyal Liaqat, Moshe Gabel, Eyal de Lara, and Tiago H. Falk. 2021. Remote COPD Severity and Exacerbation Detection Using Heart Rate and Activity Data Measured from a Wearable Device. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, Mexico, 7450–7454. https://doi.org/10.1109/EMBC46164.2021.9629949

[57] Jean-Marc Valin. 2017. A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement. (2017). https://doi.org/10.48550/ARXIV.1709.08243 Publisher: arXiv Version Number: 3.

[58] Ryhor Vashkevich and Elias Azarov. 2020. Pitch-invariant Speech Features Extraction for Voice Activity Detection. In 2020 22th International Conference on Digital Signal Processing and its Applications (DSPA). 1–4. https://doi.org/10.1109/DSPA48919.2020.9213292

[59] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 3 (Sept. 2021), 1–28. https://doi.org/10.1145/3478085

[60] Varun Viswanath, Jake Garrison, and Shwetak Patel. 2018. SpiroConfidence: Determining the Validity of Smartphone Based Spirometry Using Machine Learning. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, Honolulu, HI, 5499–5502. https://doi.org/10.1109/EMBC.2018.8513516

[61] Andrey Vyshedskiy and Raymond Murphy. 2016. Acoustic biomarkers of Chronic Obstructive Lung Disease. RIO 2 (May 2016), e9173. https://doi.org/10.3897/rio.2.e9173

[62] Tom M. A. Wilkinson, Gavin C. Donaldson, John R. Hurst, Terence A. R. Seemungal, and Jadwiga A. Wedzicha. 2004. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 169, 12 (June 2004), 1298–1303. https://doi.org/10.1164/rccm.200310-1443OC

[63] Robert Wu, Daniyal Liaqat, Eyal de Lara, Tatiana Son, Frank Rudzicz, Hisham Alshaer, Pegah Abed-Esfahani, and Andrea S Gershon. 2018. Feasibility of Using a Smartwatch to Intensively Monitor Patients With Chronic Obstructive Pulmonary Disease: Prospective Cohort Study. JMIR Mhealth Uhealth 6, 6 (June 2018), e10046. https://doi.org/10.2196/10046

[64] Wenlong Xu, Guoqiang He, Chen Pan, Dan Shen, Ning Zhang, Peirong Jiang, Feng Liu, and Jingjing Chen. 2022. A forced cough sound based pulmonary function assessment method by using machine learning. Front. Public Health 10 (Oct. 2022), 1015876. https:

//doi.org/10.3389/fpubh.2022.1015876
[65] Xuhai Xu, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. Listen2Cough: Leveraging End-to-End Deep Learning Cough Detection Model to Enhance Lung Health Assessment Using Passively Sensed Audio. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 1 (March 2021), 1–22. https://doi.org/10.1145/3448124
[66] Bendong Zhao, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu. 2017. Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics 28, 1 (2017), 162–169. https://doi.org/10.21629/JSEE.2017.01.18
[67] Yi Zhu and Tiago H. Falk. 2022. Fusion of Modulation Spectral and Spectral Features with Symptom Metadata for Improved Speech-Based Covid-19 Detection. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 8997–9001. https://doi.org/10.1109/ICASSP43922.2022.9746471

A COMPARISON TO PRIOR WORK

In Table 4, we provide a comparative overview of the literature on respiratory health sensing and speech analysis.

Table 4: Comparison of Related Work

Paper	Modality	Task/ Event of Interest	Objective	Functions without active user input	Evaluated in real-world settings	Allows direct and continuous monitoring of lung condition	Predicts pulmonary deterioration in advance
SpiroSmart [27]	Audio	Forced exhalation	Lung function estimation	No	No	No	No
SpiroCall [21], Sprio-Confidence [60], ExhaleSense [41]	Audio	Forced exhalation	Lung function estimation	No	Yes	No	No
Xu et al. [64]	Audio	Voluntary coughs	Lung function estimation	No	No	No	No
CoughGAN [43]	Audio	Voluntary Coughs	Respiratory disease detection	No	No	No	No
Nemati et al. [35]	Audio	Voluntary Coughs	Lung function estimation, disease state, disease severity	No	No	No	No
Saleheen et al. [48]	Audio	Monosyllabic voice segments	Lung function estimation	No	No	No	No

Continued on next page

Table 4: Comparison of Related Work (Continued)

Tiwari et al. [56]	IMU and PPG	Activity and heart rate	COPD symptom severity classification	Yes	Yes	Yes	No
WiBreathe [44]	Wi-Fi signals	Respiration rate	Respiration rate estimation	Yes	Yes	No	No
InstantRR [42], BreathEasy [40]	IMU	Respiration rate	Respiration rate estimation/Respiratory disease detection	Yes	No	No	No
BreathTrack [25]	Audio	Breathing Phases	Estimation of breathing phase, breathing rate, and inhalation-to-exhalation ratio; respiratory disease classification	Yes	Yes	No	No
Listen2Cough [65]	Audio	Cough	Cough detection, respiratory disease detection	Yes	Yes	No	No
Chatterjee et al. [8]	Audio	Wheeze	Wheeze detection, lung function estimation	Yes	Yes	No	No
Farrus et al. [18], Merkus et al. [30]	Audio	Read speech	Statistical analysis of correlation between acoustic features and COPD level	No	No	No	No
Chun et al. [10]	Audio	Spontaneous speech	Lung function estimation	No	No	No	No

Continued on next page

Table 4: Comparison of Related Work (Continued)

Sedaghat et al. [50]	Audio	Natural speech	Statistical analysis of correlation between acoustic features and COPD symptoms; exacerbation detection	Yes	Yes	No	No
PulmoListener	Audio	Natural speech	COPD symptom severity level classification; prediction of the onset of worsening symptoms	Yes	Yes	Yes	Yes