SpiroCall: Measuring Lung Function over a Phone Call

Mayank Goel[1], Elliot Saba[1], Maia Stiber[2], Eric Whitmire[1], Josh Fromm[1], Eric C. Larson[3], Gaetano Borriello[1], Shwetak N. Patel[1]

[1]Computer Science and Engineering, Electrical Engineering
DUB Group
University of Washington
Seattle, WA 98195

[2]Inglemoor High School
Kenmore, WA 98028
mwstiber@gmail.com

[3]Computer Science and Engineering
Southern Methodist University
Dallas, TX 75205
eclarson@lyle.smu.edu

{mayankg, sabae, emwhit, jwfromm, gaetano, shwetak}@uw.edu

ABSTRACT

Cost and accessibility have impeded the adoption of spirometers (devices that measure lung function) outside clinical settings, especially in low-resource environments. Prior work, called SpiroSmart, used a smartphone's built-in microphone as a spirometer. However, individuals in low- or middle-income countries do not typically have access to the latest smartphones. In this paper, we investigate how spirometry can be performed from *any* phone—using the standard telephony voice channel to transmit the sound of the spirometry effort. We also investigate how using a 3D printed vortex whistle can affect the accuracy of common spirometry measures and mitigate usability challenges. Our system, coined SpiroCall, was evaluated with 50 participants against two gold standard medical spirometers. We conclude that SpiroCall has an acceptable mean error with or without a whistle for performing spirometry, and advantages of each are discussed.

Author Keywords

Health sensing; spirometry; mobile phone sensing; signal processing; machine learning.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (*e.g.,* HCI): Miscellaneous.

INTRODUCTION

Portability, low-cost, and sensing capabilities provide mobile phones a distinct advantage in health sensing. Phone-based health applications often save patients from using and carrying dedicated medical devices. This advantage is particularly apparent in the management of chronic diseases, where patients frequently use health tests to monitor disease progression and manage treatment.

Figure 1. A user using SpiroCall on a feature phone (Sony w580i) with, and without a SpiroCall whistle.

Recently, a number of health applications have been developed to estimate physiological measures such as heart rate [13], respiratory rate [8,12], pupillary dilation [14], and newborn jaundice [5]. Larson *et al.* [9] introduced SpiroSmart, a smartphone-based spirometer that measures a user's lung function using the phone's built-in microphone. Spirometry is the mainstay for measuring lung function and standard of care for diagnosing chronic lung impairments, such as asthma, chronic obstructive pulmonary disease (COPD), and cystic fibrosis. During spirometry tests, participants forcefully exhale as much air as they can from their lungs. The spirometer measures the instantaneous flow and cumulative volume of exhaled air. It then calculates multiple lung function measures to help diagnose and manage various pulmonary conditions.

Introduced in 2012, SpiroSmart [9] is a smartphone application that records the user's exhalation and sends the audio data generated to a central server. The server then calculates the expiratory flow rate using a physiological model of the vocal tract and a model of the reverberation of sound around the user's head. SpiroSmart was an important step in making spirometry more accessible, and since its introduction, it has been involved in numerous clinical

studies. SpiroSmart is currently deployed in multiple locations around the world, including Seattle and Tacoma in USA, Khulna in Bangladesh, and Pune in India. Thus far, we have collected data for around two thousand patients using SpiroSmart with encouraging results. While an analysis of the collected data is not the focus of this paper, we highlight four challenges that have surfaced from the SpiroSmart deployments: (1) SpiroSmart requires a smartphone; (2) usability and training challenges exist; (3) a patient with severely low lung function might not generate any sound; and (4) algorithms created from audio collected on a specific smartphone model may not generalize to other models or brands. In this paper, we critically examine ways to address these challenges and evaluate our proposed solutions with a set of 50 new patients.

Smartphones are becoming prevalent at a breathtaking rate, yet more than half of the mobile phone users in sub-Saharan Africa and South Asia will still be using a non-smartphone (or feature phone) in 2020 [4]. A major portion of the population suffering from lung impairments lives in these low resource environments. In fact, according to a recent WHO report, more than 90% COPD deaths occur in low- and middle-income countries [19]. Thus, we believe that phone-based spirometers need to work on all mobile phones, and not just programmable smartphones. Even smartphones, the diversity of phone manufacturers and models makes it challenging to manage custom applications for every type of mobile phone.

To this end, we present *SpiroCall* (Figure 1), a call-in service that measures lung function on *any* mobile phone without the need for a locally running application. Unlike SpiroSmart, it transmits the collected audio using the standard voice telephony channel. A server receives the data of degraded audio quality and calculates clinically relevant lung function measures and reports to the participants using audio or text message. The ability to use a server to analyze audio data transmitted from any mobile phone, be it a feature phone or smartphone, eliminates the need to develop a specialized application for every phone platform. SpiroCall combines multiple regression algorithms to provide reliable lung function estimates despite the degraded audio quality over a voice communication channel.

Although the call-in service removes the need for a smartphone, there are other significant usability challenges that are more difficult to mitigate: how a user holds the phone (angle, microphone occlusion, *etc*.), the distance from the user's mouth to the phone, and how wide a user opens their mouth. Recognizing that some people may not be able to master the technique needed to perform this maneuver, we also designed a simple and low-cost 3D-printed whistle accessory. The whistle (Figure 1, *Top*) generates vortices as the user exhales through it [17,18], changing its resonating pitch in proportion to the flow rate. The whistle does not have any moving parts and is as simple as any spirometer mouthpiece. Despite the additional hardware, the whistle offers several important advantages: (1) the acoustic properties of the whistle are more consistent than a user's vocal tract and generate audible sounds even at lower flow rates, (2) the whistle removes the effect of distance from the user's mouth, and (3) precisely controlling mouth shape and phone orientation are less important. In this paper, we investigate viability of the call-in service approach with and without the whistle.

We evaluated SpiroCall in a controlled study with 50 patients. We compare SpiroCall to two FDA approved spirometers and evaluate the effect of using the voice communication channel on the performance of SpiroCall. Each patient performed spirometry efforts with and without the whistle on two different phones recording the audio through the cell phone network and two smartphones recording the audio locally through an app. Participants used two different sizes of vortex whistles to determine whether different sizes work better for different individuals. Our results show that without a whistle, SpiroCall has a mean error of 7.2% for the four major clinically relevant lung function measures. For FEV1% (the most commonly used diagnostic measure [2]), the mean error is 6.2%. With a whistle, SpiroCall has a mean error of 8.3% for the four measures, and 7.3% for FEV1%. Although, using the whistle leads to higher average error in lung function estimation, it performs more consistently for people with lower lung function and produces fewer over-estimations of lung function (*i.e.*, false negatives), as compared to when not using a whistle.

The main contribution of this paper is a demonstration that *every* mobile phone in the world can be used as a spirometer. This contribution comes in four parts: (1) an algorithm to estimate lung function from a standard telephony voice channel's degraded audio signal; (2) a custom-designed whistle that reduces usability and performance challenges; (3) a comparison of the call-in service and the whistle against two clinical spirometers (using different phones); and (4) a demonstration of how poor quality audio, transmitted across the standard telephony voice channel, can be utilized for modeling and inference.

BACKGROUND OF SPIROMETRY

Spirometry is the most widely employed pulmonary function test. Many different types of spirometers are available, ranging from big, clinical spirometers to portable, home spirometers. Their cost also varies from $1,000 USD to $5,000 USD. During a spirometry test, the patient takes the deepest breath possible and then exhales with maximum force for as long as possible. The spirometer measures the amount and speed of airflow and calculates various lung function measures based on the test. Four of the most important lung function measures are:

(1) Forced Vital Capacity (FVC): Total volume of air expelled during the expiration,

(2) Forced Expiratory Volume in one second (FEV_1): Volume of air expelled in the first second of expiration,

(3) FEV_1/FVC (FEV1%): Ratio of FEV_1 and FVC, and

(4) Peak Expiratory Flow (PEF): Maximum expiratory flow rate reached during the test.

A healthy individual's lung function measures are generally at least 80% of the values predicted based on their age, height, and gender [7]. Abnormal values of FEV1% are (expressed as a percent of predicted value) [10]:

- Mild to Medium Lung Dysfunction: 60-79%,
- Moderate Lung Dysfunction: 40-59%, and
- Severe Lung Dysfunction: below 40%.

Apart from the numerical measures, spirometers also generate flow *vs.* time, flow *vs.* volume, and volume *vs.* time plots. Figure 2 shows examples of FV plots. In a healthy individual, the descending limb of the FV plot is almost a straight line (black, solid line in Figure 2). As obstruction to the airflow increases, the flow rate decreases faster than exponentially after reaching its maximum value (PEF). Therefore, it attains a curved or "scooped" slope (blue, dashed line in Figure 2). For an individual suffering from a restrictive lung disease, such as cystic fibrosis, the respiratory muscles weaken and the patient's lung capacity (FVC) decreases (red, dashed line in Figure 2).

DESIGN OF SPIROCALL

The previous work of SpiroSmart offloaded a significant chunk of computation to a server, with the audio transferred via an Internet connection. Thus, the received audio was lossless and free of artifacts. In SpiroCall, we leverage the voice communication channel to transmit the audio data to a server. The server then uses machine learning to compute lung function measures. The features used by our machine learning model fall into three categories: temporal envelope detection, spectrogram processing, and linear predictive coding (LPC). The cellphone channel (GSM) uses LPC to encode voice. This means that even though the GSM channel compresses the audio signal, the values of LPC coefficients remain largely preserved. Landlines, or POTS (Plain Old Telephone Service) is also an attractive option as a communication channel. However, we do not focus on POTS in this paper because GSM networks are far more prevalent than landlines in the developing world.

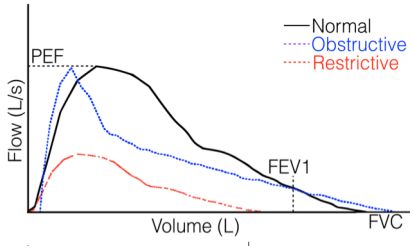

Figure 2. Example of different Flow *vs.* Volume curves and major lung function measures.

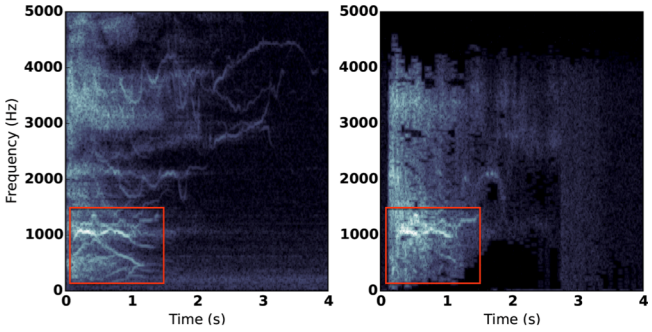

Figure 3. *(Left)* Spectrogram of a spirometry effort recorded locally, and *(Right)* recorded through the voice channel. The GSM network downsamples the audio, and the data over 4 kHz is lost. However, the data in our main region of interest (red square) is largely reconstructable.

Lung Function Estimate without Whistle

Figure 3 shows spectrograms of a spirometry effort recorded locally on a smartphone *(Left)* and the same effort after the data is sent through a GSM cell phone network *(Right)*. The device type is an iPhone 4S in both cases. There are significant differences between the two spectrograms as the voice undergoes many changes as it goes through a communication channel. Although different communication networks use different speech coding techniques, all GSM/UMTS speech-coding algorithms share similarities in their treatment of speech and are based upon the same underlying linear prediction approach.

First, all GSM voice coding technologies use a source-filter model for speech. That is, the "source" estimates the lung or glottis excitation, and the "filter" estimates how the vocal tract blurs this excitation into continuous sound. Parameters of the source and filter are then transmitted through the channel, instead of the raw audio. The most common method for separating out the source excitation from the vocal tract filter is to use LPC. An artifact of the LPC calculation is that the strong frequency resonances are preserved (and are calculable directly from the LPC coefficients). These resonances are also the primary features in our algorithms. As such, we expect the LPC encoding to preserve much of the important information in the signal. An example of this is shown in Figure 3 – the fundamental resonance is easily seen in both recordings (inside the red box), despite many smaller details, such as higher harmonics of the fundamental resonance and all spectral energy above 4 kHz, being lost.

Additionally, the transmission process suppresses low-energy components in the signal, as can be seen in Figure 3 *(Right)* where the energy of the signal abruptly cuts off in patches. In contrast, the signal maintains high fidelity and stays above the noise floor for the initial (and relatively louder) segment of the effort (inside the red box in Figure 3, *Right*).

Algorithm for Lung Function Estimation over a Voice Channel without a Vortex Whistle

In order to deal with the drastic variation in sound quality as the data goes through a GSM channel, we sought to evaluate what modifications are necessary to the algorithm proposed in the original SpiroSmart paper [9].

We use the microphone as an uncalibrated pressure sensor and the received pressure values are transformed using three approaches (Figure 4): (1) envelope detection, (2) resonance tracking in the frequency domain, and (3) linear predictive coding (LPC). The envelope of the signal can be assumed to be a reasonable approximation of the flow rate because it is a measure of the overall signal power (or amplitude) at low frequencies. In the frequency domain, resonances can be assumed to be amplitudes excited by reflections in the vocal tract and mouth opening—and therefore should be proportional to the flow rate that causes them. Finally, we can use linear prediction as a flow approximation. Linear prediction assumes that a signal can be divided into a source and a shaping filter and it estimates the source power and shaping filter coefficients. The "filter" in our case approximates the vocal tract. The "source variance" is an estimate of the white noise process exciting the vocal tract filter—in our case, this approximates the power of the flow rate from the lungs. Each approach generates multiple time-domain flow-rate estimations.

We extract separate feature sets for FEV1, FVC, and PEF from these time-domain flow-rate estimations. For example, PEF is defined as the maximum flow reached in an effort. Thus, for a given flow-rate estimation, we take the max value and use it as a feature for PEF regression. In contrast, FVC is defined as the total volume of air exhaled. Thus, integrating the flow-rate estimation with respect to time gives us a feature for FVC regression. Using this approach, we generate 3 sets of of 38 features for FEV1, FVC, and PEF, each. We do not use any regression algorithm to estimate FEV1%. This value is simply a ratio of the estimated FEV1 and FVC.

Considering that the GSM channel uses LPC to encode sound, the LPC-based features used in our algorithms remain largely preserved. The envelope detection-based features are based on the coarse amplitude of sound with respect to time; in most cases these features remain preserved as well. The spectral features are most affected by the GSM channel because the high frequency details are completely lost. However, upon analysis we realized that the resonances within the first harmonics were strong enough that most spectral features contain some relevant information.

In the original algorithm, the calculated features were sent to a random forest regression that, because it has no underlying linear model, had trouble exploiting some of the linearity in the feature data. The algorithm performed poorly on the data collected through the GSM channel and

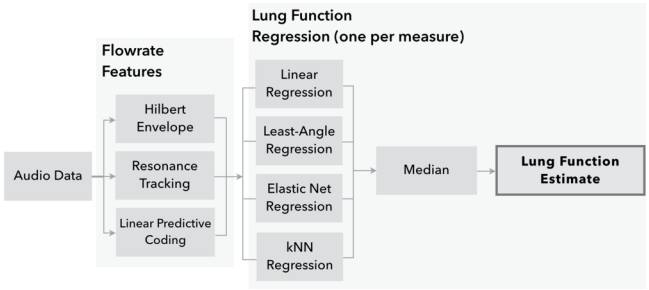

Figure 4. Flowchart for lung function estimation without whistle.

it over-estimated the lung function for participants with obstructed lungs (FEV1% < 0.8). Therefore, we have updated the algorithm to employ an ensemble of four different regression algorithms (Figure 4), with the aim that each regression would provide a different perspective. We run the regressions using the *scikit-learn* toolkit in Python and use leave-one-patient-out cross-validation to avoid overfitting. Furthermore, we keep all the parameters for all the algorithms at their default values and do not tune them for the collected data.

The first regression is a linear regression that tries to find a linear relationship between the features and the ground truth lung function value. The second regression uses least angle regression (LARS) [3]. LARS selects the most useful features using a variant of forward feature selection, but the underlying model is assumed to be linear. The third regression uses the elastic net algorithm [20], which eliminates features in a slightly different way than LARS. This regression uses a combination of LASSO regression and ridge regression for regularization that is often more stable. Finally we use enclosing k-Nearest Neighbor regression (k = 2) [6], which finds the convex hull of the data in the feature space and fits a locally linear regression. Though the underlying model is assumed linear, the local fitting often can fit many different types of nonlinearity. We find the final regression estimate by taking the median of these four regressions. We use this same process for FEV1, FVC, and PEF measures. As mentioned, FEV1% is calculated as a ratio of the estimated FEV1 and FVC values.

Additionally, there are situations when the test is performed in a noisy environment or the channel itself might be noisy. To deal with such situations, the system automatically detects the level of background noise by looking at the mean absolute amplitude of the recorded sound for a 250 ms window immediately before the user exhales. This is the period when the user is most silent and we use it as an opportunity to measure the ambient noise level. If the amplitude of sound within this window is estimated to be above an empirically determined threshold, the environment is considered unsuitable for data collection. The threshold used here is same as the one used in the SpiroSmart clinical trials.

Lung Function Estimate with Whistle

SpiroCall faces three audio sensing challenges, (1) variability of different phones, (2) low sound amplitude for severely impaired patients, and (3) inconsistency of the distance between a user's mouth and the microphone. Bernard Vonnegut, in 1954, designed a whistle that changed its pitch in proportion to flow rate and called it a vortex whistle [17]. Later, Watanabe and Sato suggested modifications to vortex whistle construction for use in spirometry efforts [16,18,21]. In their study, they used pitch tracking to convert the vortex whistle sound to an estimate of flow rate. Considering pitch tracking is resilient to variations across devices, such as gain and frequency response, the whistle could make SpiroCall independent of distance, channel, and device. The whistle has no moving parts and thus, is as simple as any spirometer mouthpiece—mass-producible for less than 10 cents (US). We decided to test the design proposed in [16] to see if it could be used as a flow-sound transducer instead of the user's vocal tract. However, we found the proposed design unsuitable for spirometry and modified it based on a pilot study with 15 participants.

The vortex whistle consists of three sections: the inlet, the cylindrical cavity, and the downstream tube. The *inlet* is a cylindrical pipe that is tangentially connected to the cylindrical cavity on its curved surface. The user blows through this tube. The *cylindrical cavity* allows the air inside to swirl around the chamber. The *downstream tube* is attached perpendicular to the cylindrical cavity. When the air enters the cylindrical cavity, it starts rotating along the circumference of the cavity, thereby forming a vortex, and moves toward the downstream tube. The arrows in Figure 5 *(Left)* show the result of a simulation of airflow within the whistle. The color of the arrow denotes the simulated velocity of the air. When the air leaves the cylindrical cavity, the vortex becomes unstable and whips around at an angular velocity that is proportional to the rotational velocity of the vortex. This unstable vortex generates sound as it leaves the downstream tube.

The frequency produced by the whistle is affected by several factors, including the dimensions of the whistle [17]:

$$F = \frac{U}{2\pi R_{CC} A} sin\ \theta \sqrt{\frac{1}{R_f(L_{DST}+\Delta L_{DST})}} \qquad (1)$$

where F is frequency, U is input flow rate, R_{CC} is the radius of the vortex in the cylindrical cavity, A is the cross-sectional area of the inlet, R_f is radius of the air in the downstream tube, L_{DST} is the length of the downstream tube, and ΔL_{DST} is the length of the vortex formed at the outlet of the whistle. The *sin* term refers to the angle between the formed vortex and cylindrical plane. This term is difficult to calculate mathematically, necessitating that the quantity be determined through calibration [16]. Typical values range between 0.35 up to 0.95 [16].

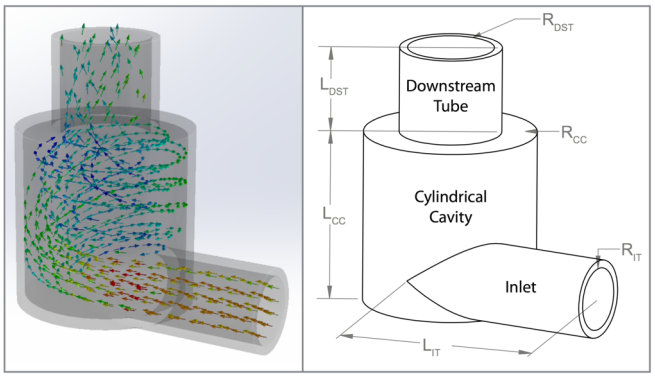

Figure 5. *(Left)* 3D rendering of the whistle. The arrows show the airflow and the colors denote the velocity. Red is faster and blue is slower. *(Right)* Dimensions (in mm) of the two whistles: Small: R_{CC} = 20, R_{DST} = 11, R_{IT} = 8, L_{IT} = 50, L_{DST} = 24, L_{CC} = 40. Big: R_{CC} = 37.5, R_{DST} = 12.5, R_{IT} = 8, L_{IT} = 74, L_{DST} = 27, L_{CC} = 35.

We modified the design suggested in [16] to ensure that the whistle's response remains linear even at flow rates around 15 L/s (verified via SolidWorks™ simulations). This flow rate is well above the peak flow rate attainable by individuals with height up to 210 cm. We designed two sizes of the whistle (dimensions shown in Figure 5, *Right*), as different sizes will have different pitch gradients.

We 3D-printed the whistles on a Stratasys BST768 printer using ABS plastic material. Our evaluation of both the whistle sizes with 50 participants demonstrated that the bigger whistle performed better because it had a steeper pitch gradient. From a usability standpoint, 34 out of 50 participants also preferred using the bigger whistle, because it was easier to handle.

Algorithm for Lung Function Estimation with Whistle

When a vortex whistle is used, we can simplify the audio processing considerably because the whistle pitch changes linearly in response to flow rate. Simple pitch tracking can estimate the flow rate over time. We can calibrate the parameters of this linear relationship (bias and slope) using a few example spirometry efforts. For a particular vortex whistle with set dimensions, these parameters only need to be calibrated once.

Whistle Pitch Extraction: All audio data is resampled to 44.1 kHz to ensure uniformity in the processing across devices with different sampling rates. We first process the spectrogram of the effort to track the pitch. We segment the data into frame durations of 46 ms with a step size of 3 ms between frames. Next, we find the peak magnitude in the spectrogram (Figure 6) and search for the peak frequency within 0.25 seconds. The peak frequency (Figure 6, top of the white curve) corresponds to the PEF of the spirometry effort. We track pitch backward and forward in time from that point, stopping when the spectral energy ceases to trend towards lower frequencies. This helps us ignore wheezing at the end of a spirometry effort that may overwhelm the

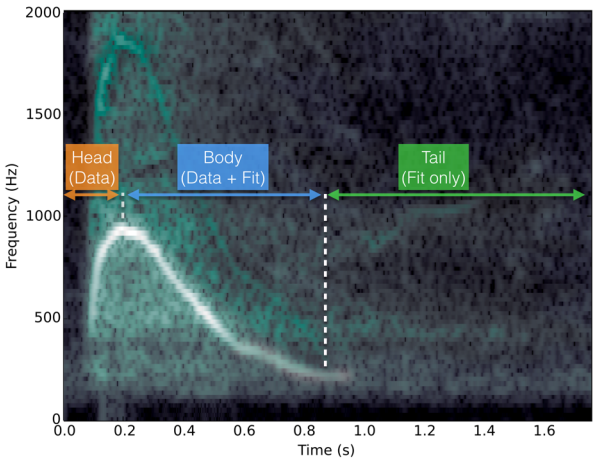

Figure 6. The blue and orange regions are associated with the pitch tracked by the algorithm. The green region is extrapolated based on information in the blue region.

main whistle audio amplitude. For each frame, we fit a quadratic polynomial to the frequency bin of interest and its two neighbors to attain sub-bin accuracy in our peak frequency estimates [11]. We stop pitch tracking once the resonance passes below a certain empirically determined pitch threshold, as the whistle mouthpiece does not resonate well at lower flow rates and therefore lower frequencies.

For example, in Figure 6, the pitch can be tracked up to 1 s. This means that while we can infer the FEV_1 value from the pitch data, the FVC value needs an extrapolated curve.

Tail Extrapolation: After the flow achieves its peak value (PEF), the flow rate decays exponentially for a healthy individual and decays faster than exponentially for an individual with obstructive lung impairments. Therefore, when we extrapolate the pitch curve, we cannot just use an exponential fit function. We apply a combination of exponential and exponential of exponential fits, so that the system automatically adapts to different types of flow-time curves, including the ones where the flow rate decay faster

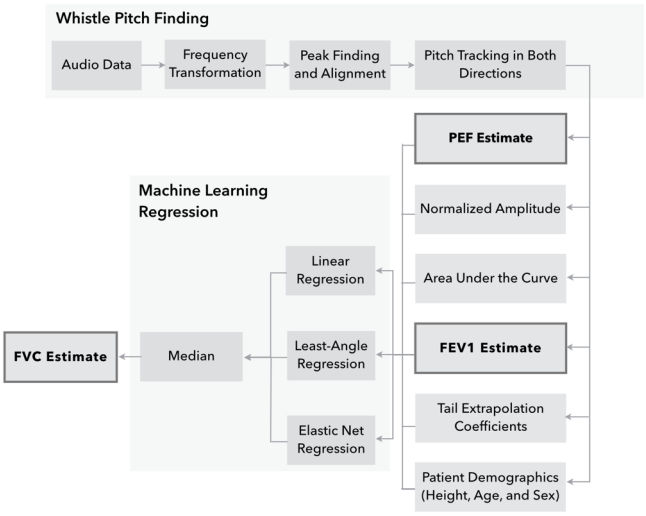

Figure 7. Flowchart for lung function estimation with whistle.

than exponentially . We fit the following function to the tail end of the flow-time curve:

$$x(t) = \left(a_0 e^{-b_0 t} + a_1 e^{-b_1 t^2}\right) \cdot a_2 e^{e^{-b_3 t}} \qquad (2)$$

We use the entire descending limb of the tracked pitch to fit our extrapolation function. The green area (Tail Fit only) in Figure 6 shows the time over which the curve is extrapolated. The blue area (Body, Data+Fit) represents the phase in time during which reliable resonance tracking data is available. However in order to transition smoothly from resonance-tracking to extrapolation, we cross-fade from resonance-tracked data to extrapolated data within the blue region. We evaluated our extrapolation function by applying it to the set of groundtruth flow-time curves to ensure it was able to model the tail end of a user's exhalation. Although the extrapolation function worked exceptionally on groundtruth data (mean error = 3.2%), when we evaluated the extrapolation on the audio data received from SpiroCall devices, our FVC estimates had an average error of 15%. We therefore decided to estimate the FVC through a regression model, using the extrapolated curve as a feature in the regression.

FVC Regression Model: Although our tail extrapolation method did not provide an adequate volume (FVC) measure, it still provided a good, albeit noisy, estimate in most cases. We, therefore, encode the pitch tracking output as a set of regression features. Figure 7 shows all the features used in the regressions. The features can be broken down into the three phases of the pitch tracking in Figure 6: *Head, Body,* and *Tail.* We use the estimated PEF, *i.e.,* the peak frequency of the tracked pitch, as the representative feature from the *Head* section of the curve. We also use the peak amplitude (normalized) of the overall audio. From the *Body* section of the curve, we use the area under the pitch-tracking curve until the end of the body section, and the area under the curve until the end of 1 sec, *i.e.,* FEV_1 estimate. The next set of features comes from the *Tail* extrapolation. We use the coefficients generated by the tail extrapolation as an encoding of the curve in the *Tail* region of the curve. Specifically, we use a_0, a_1, and a_2 from Equation (2) as features in the regression. Apart from these features, we also use height, age, and sex as our demographic features. It is common practice in spirometers to record a patient's physical details as this information helps the device in calculating predicted normal lung function for the patient.

Similar to the no-whistle condition, the regression algorithm employs an ensemble of three regressions: linear, LARS, and elastic net regressions. We combine the outputs of all regressions and select a median of their estimates as the final FVC estimate. We use leave-one-patient-out cross validation in all levels of learning to avoid overfitting.

EVALUATION

To evaluate SpiroCall, we created an extensive dataset of audio samples and ground truth spirometry data. We recruited 50 participants (30 males, 20 females), ranging in age from 21 to 67 years (M = 30) through flyers and email messages in the university. The study sessions were conducted in a non-clinical lab setting and lasted for approximately 30 minutes. 20% of participants had mild to moderate lung obstruction, *i.e.,* FEV1% < 0.80 (Table 1).

The SpiroCall study used a within-subjects $2 \times 2 \times 3$ factorial design. The factors and levels were:

- ***Phone Type:*** *iPhone* and *non-iPhone.* We used two non-iPhone devices: Samsung Note 3 and Sony Ericsson W580i. We used the W580i (feature phone) to evaluate the performance of SpiroCall on an approximately 10-year-old device.
- ***Channel Type:*** *Local recording* or *voice channel recording.* We kept the iPhone consistent in both channels to analyze the performance of SpiroCall if only the channel is changed.
- ***Whistle:*** *No whistle, small whistle,* and *big whistle.* We recorded audio data for two whistles to understand if different participants preferred different sizes or if one size gave results that were more reliable that the other.

All the conditions were counterbalanced and we randomized the order of the whistles.

Experimental Setup

We collected the audio data on four phones, two iPhone 4S smartphones, a Samsung Galaxy Note 3, and a Sony Ericsson W580i feature phone. All four phones were in front of the user at roughly an arm's length away (Figure 8). The distance was not formally controlled or varied. One of the iPhones and the Samsung Note recorded the audio data locally at 32 kHz and 44.1 kHz, respectively. The other two devices sent the data over the GSM voice channel. These phones placed phone calls to different Google Voice accounts that recorded the data in the form of voicemail messages. Google Voice saved the audio data as 44.1 kHz MP3 files, but the GSM channel band-limited the data to less than 8 kHz. The difference between the local recordings and those done over the GSM channel is shown in Figure 3.

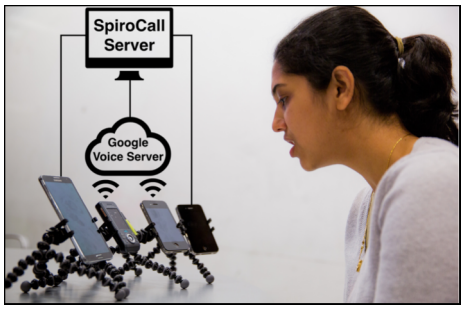

Figure 8. SpiroCall experimental setup. We recorded the data on four phones at the same time. Two phones recorded the audio locally. The other two phones called Google Voice numbers and sent the audio data over the GSM channel to the Google Voice server.

Table 1. Demographic information of the participants.

Participant Demographics (N = 50)	
Males (n, %)	30 (60%)
Age (yrs) (mean, range)	30 (21 – 67)
Height (cm) (mean, range)	172 (155 – 188)
Reported Lung Ailments Asthma: 10 (20%), Bronchitis: 2 (4%), COPD: 2 (4%), Cystic Fibrosis: 1 (2%), Sarcoidisis: 1 (2%)	
Low Lung Function (n, %)	16 (32%)
Never Performed Spirometry (n, %)	30 (60%)

We transferred the data from the local phones (iPhone 4S and Samsung Note) to the computer over a USB connection at the end of the study. We downloaded the data from the Google Voice accounts as MP3 files to a computer.

Procedure

We collected the ground truth for the participants on two FDA-approved clinical spirometers: the nSpire Koko Legend and the NDD EasyWare spirometer. We used the two spirometers to answer two questions: (1) whether the participants got fatigued as the session progressed, and (2) how much variability exists between the outputs of the two devices. We recorded the variability between the clinical devices to use it as a benchmark for SpiroCall's performance. The participants performed at least 15 spirometry efforts (three each for: two clinical spirometers, two whistles, and one without whistle). Spirometry measurements are completely effort-dependent and some fatigue can build up when performing this many efforts. Therefore, we recorded efforts on one clinical spirometer at the beginning of the session and on another spirometer at the end of the session. We randomized the order for each participant.

At the start of each session, we explained the forced expiratory maneuver to the participants and we asked them to practice using the spirometer. Once the participants were able to perform an acceptable maneuver according to the ATS criteria for reproducibility [10], three efforts were recorded using the spirometer. Next, we introduced the participants to SpiroCall.

The four phones (Phone Type × Channel Type) recorded the audio simultaneously, thus saving the participants from performing tests with each device type separately. One of the authors, who was trained to administer spirometry efforts, gave feedback to the participants regarding the acceptability and quality of the efforts. In the future, it will

be straightforward to have a system that automatically determines if an effort was too low in volume.

Note that collecting the SpiroCall data and the clinical spirometer data at the same time is impossible, so explicit ground truth was unknown. Instead, each effort from SpiroCall was associated with the best effort selected by the clinical spirometer. As per the ATS criteria, the spirometer selects the effort with the highest FVC as the best one [2,10].

RESULTS

In this section, we discuss the performance of SpiroCall when compared to the two clinical spirometers in terms of accuracy of estimated lung function measures and false positives *vs.* false negatives. We consider an estimate to be a false negative if the groundtruth FEV1% is below 0.8 and SpiroCall predicts the value to be above 0.8 [2]. We break down these results by *Phone Type* and *Channel Type.* We also compare the performance of SpiroCall with and without a vortex whistle. Finally, we discuss the accuracy and usefulness of the flow-volume curves generated by SpiroCall. Based on our evaluation we conclude that SpiroCall can help in screening and monitoring patients with lung impairments in low resource regions.

Two Ground Truth Devices

As mentioned, we used two clinical spirometers to collect groundtruth. We compared their respective lung function measure and found that PEF had the maximum difference of 9.2% between the two devices, and FEV1, FVC, and FEV1% had a difference of 5.1%, 5.2%, and 3.2%, respectively. However, none of these differences are statistically significant (based on an F-test, $p>0.05$). We also studied the effect of order to understand if fatigue played any role in exaggerating the difference between the two devices. In a 2-way ANOVA test with *presentation order* of the two spirometers as a between-subjects factor, we found that the difference in estimates of PEF and FEV1 were statistically significant ($p<0.05$). This finding suggests that the participants got fatigued by the time the session ended. Therefore, we use the results from the *first* spirometer that the participants used as their groundtruth or reference device. While this means that the reference device was not consistent across participants, the difference in device performance was not found to be significant and should not strongly affect the final analysis. In addition, we corrected for fatigue by counter-balancing between all *Phone, Channel*, and *Whistle Type* conditions for all participants.

Lung Function Estimate without Whistle

We break down the comparison of measurements from SpiroCall and the clinical spirometers by evaluating how well it performs for different lung function measures and the number of outliers and false negatives.

Accuracy of Lung Function Measures

The graphs in Figure 9 *(Left*) present the percentage error of each measure without a whistle. For all lung function measures, the algorithm returns an average error of less than 10%. There is no significant difference between the performance of smartphones recording the data locally in an app (Samsung Note and Apple iPhone) and phones running over the voice communication channel (Sony W580i and Apple iPhone 4S). The performance is best for FEV1%, which is the most common measure of lung function used in diagnosis because it is typically most consistent [10,15]. The mean error rate for FEV1% is below 6% for all the four conditions. The ATS acceptability criteria require lung function measures to be within 7% to 10% of one another [15]. For most patients, SpiroCall performs well within the expected level of variation, even if the patient did not have a smartphone and performed the test on a phone call. However, it is important to evaluate the outliers (with error higher than twice the standard deviation) and see whether the lung function measures are under-estimated or over-estimated. We use twice the standard deviation because the first standard deviation is within the ATS acceptability criteria [15] and the result cannot be considered an outlier.

Outliers and Patients with Low Lung Function

In order to understand the direction of the bias, we use the modified Bland-Altman plots [1] in Figure 10. The figure shows the percentage difference between SpiroCall and the output of a spirometer *versus* the spirometer measurement of FEV1%. Lines indicating $\pm 2\sigma$ (red dashes). We focus solely on FEV1% because it is the most common lung function measure for diagnosis. If the percentage difference is positive, then the lung function was over-estimated (false negative). It can be seen in Figure 10, *Top-Left* and *Top-Right,* that SpiroCall (both without whistle) tend to over-

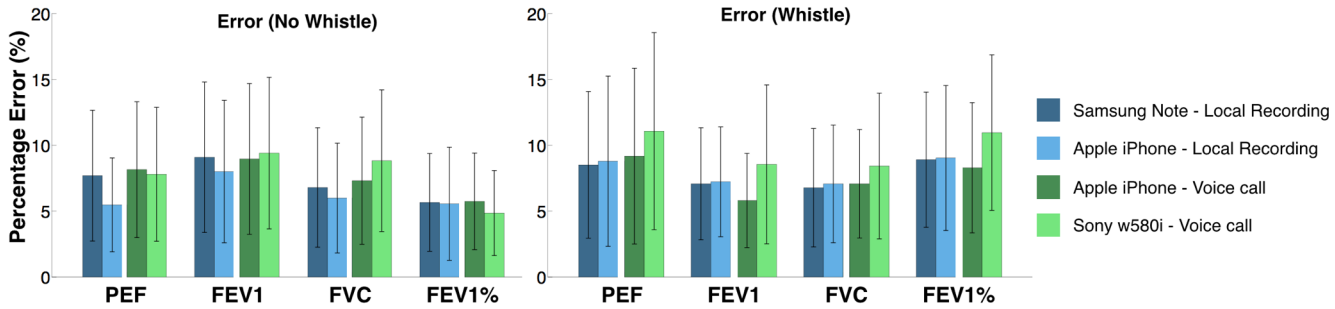

Figure 9. Percent error for different lung function measures on different devices without whistle (Left), and with whistle (Right). The first two devices recorded the data locally in an app; the next two devices recorded the data over a phone call. The error bars show standard deviation.

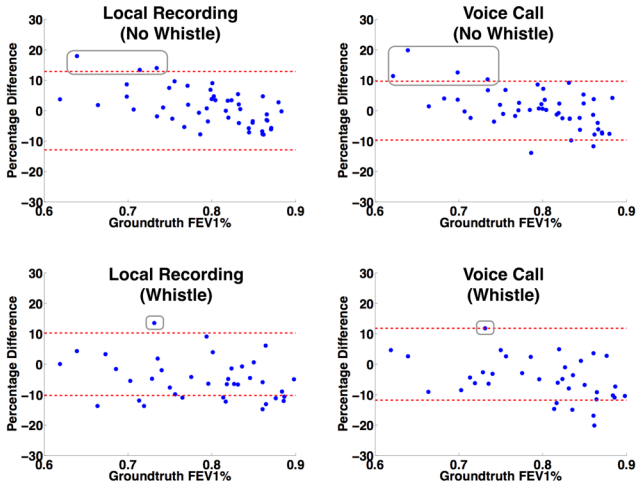

Figure 10. Bland-Altman plots of percent error of FEV1% (without and with whistle) for local and voice call recordings versus the value obtained from the clinical spirometer. The false negatives are highlighted inside grey boxes. ±2σ (red dashes) are also shown.

estimate the actual value for some patients with low lung function (FEV1% < 0.8), *i.e.,* a false negative. We highlight the false negatives inside the gray boxes. In any medical device, it is more acceptable to have a false positive than a false negative. The main reason the system currently has more false negatives for low lung function is because the algorithm is data driven and the population with higher lung function is better represented. Therefore, the model tends to bias towards the median value. Considering the signal-to-noise ratio is lower for the devices connected over the GSM channel, the false negatives are slightly more pronounced in case of voice call (Figure 10, *Top Right*).

One way to quantify the model's bias towards higher lung function is to calculate the statistical effect of lung function measure (FEV1% in this case) on the error of the model. We tested for effects of groundtruth FEV1% on the percent error through a chi-square test. We found that there was a significant effect of the groundtruth FEV1% on the accuracy of SpiroCall ($p<0.05$). As such, the performance of SpiroCall might degrade further if tested on more highly obstructed patients. Although the bias is only slight and there are relatively few false negatives, from a diagnostic perspective, it could mean that patients are screened improperly.

Lung Function Estimate with Whistle

The bias in the performance of the system due to groundtruth lung function of the user prompted us to explore the possibility of using a whistle for the users of SpiroCall.

Comparison of Two Whistle Sizes

We used two sizes of the vortex whistle in our study. Both whistles had slightly different gradient of pitch with respect to the input flow. We performed a two-sample F-test for equal variances on the percentage error for the four lung function measures for both whistle sizes. We observed a significant effect ($p<0.01$) of size on FVC and FEV1%, in favor of the bigger whistle. The percentage difference between the two whistles was 0.24%, 4.22%, 2%, and 2.13% for PEF, FEV1, FVC, and FEV1%, respectively. Considering the bigger whistle worked significantly better, our analysis of SpiroCall only includes the larger whistle.

Accuracy of Lung Function Measures

Bar graphs shown in Figure 9 *(Right)* display the percentage error of each lung function measure for each device and connection type with a whistle. The Sony Ericsson W580i performed the worst among all the phones. However, the difference was not statistically significant (F-test, $p>0.05$). Among the lung functions, the error was highest for PEF, but it is worthwhile to note that the variance in PEF was also the highest for the groundtruth spirometers. The most widely used lung function measure, FEV1%, has less than 8% mean error for three of the four device types.

Outliers and Patients with Low Lung Function

In order to understand the direction of the bias present in whistle results, Figure 10 *(Bottom)* shows modified Bland-Altman plots of FEV1%, displaying percentage difference between SpiroCall (with whistle) and the spirometer *versus* the spirometer measure. From these plots, we show that the whistle mitigates false negatives. We highlight the false negatives inside gray boxes. Most of the error for the whistle comes from false positives. When comparing local recordings and voice calls, there is no significant performance difference (F-test, $p>0.05$). However, using the whistle, we eliminate the bias in the estimate that we saw in case of no whistle. This means the whistle may be a superior screening tool, especially for patients with very low lung function. We quantify this effect of bias as before by considering the effects of groundtruth FEV1% on the percent error through a chi-square test. We found that there was *no significant effect* of the groundtruth FEV1% on the accuracy of SpiroCall across devices ($p>0.05$).

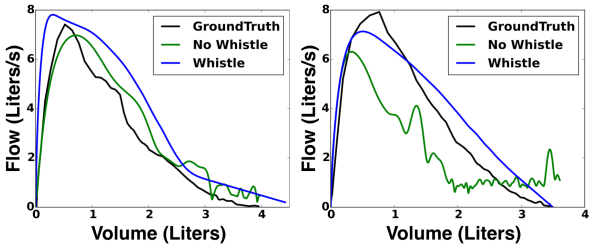

Figure 11. Two Flow *vs.* Volume curves generated by SpiroCall with a whistle and without a whistle.

Curves Generated by SpiroCall

We now shift our discussion from lung function measures to the shape of the flow-volume curves. The spirometry curves serve two purposes: (1) to evaluate if the patients performed the effort sufficiently, and (2) to help in diagnosis by showing the descending limb of the FV curve. A technician looks at the slope of the FV curve from the

start of the test to PEF. This slope should be as steep as possible, indicating that the initial blast of air was truly maximal. The investigator also looks to see if the user coughs during the spirometry maneuver. Coughing makes the descending edge of the FV curve non-monotonic as the user ends up inhaling during a cough. Therefore, it is important to evaluate how SpiroCall performs in generating these curves.

Figure 11 shows example flow-volume curves generated by SpiroCall without the whistle and with the whistle; we find that the curves generated without a whistle can be unreliable. The no-whistle (green) curve in the Figure 12 *(Right)* has an inaccurate shape because the latter half of the effort by the patient was very quiet. When the GSM channel compressed the audio, this segment was heavily compressed and not reconstructed accurately. However, these curves can still be used for validity assessment of the efforts. The initial part of the effort is always very loud and reconstructed accurately. Therefore, the investigator can still look at the ascending slope at the start of the test. For cough information, we envision that the Hilbert envelopes of the temporal audio data can be attached along with the spirometry curves, which would make any coughs clearly visible. However, in cases where the spirometry curves are of importance, we suggest the use of a whistle. The whistle generates a direct mapping to the Flow *vs.* Time curve and the final Flow *vs.* Volume curves are usually very accurate. We recognize that a more rigorous evaluation of the spirometry curves is important. This is part of our on-going work, where we are sending all the curves generated by SpiroCall to medical practitioners for quality assessment at Spirometry 360[1].

DISCUSSION

SpiroCall offers two approaches to performing spirometry through a call-in service: with a vortex whistle and without. The performance of both approaches is very promising and the mean error of the four major lung function measures is 6.2%, which is well within the ATS criteria for a clinical spirometer. However, the system sometimes over-estimates lung function when used without a whistle. We believe that this limitation stems from the fact that without a whistle, the algorithm depends on the spread and variation in its training data to remove the bias in its estimation. We plan to combine the SpiroCall clinical evaluation with ongoing SpiroSmart clinical trials.

The linear relationship between flow-rate and pitch makes the vortex whistle reliable for estimating lung function measures and spirometry curves with significantly fewer false negatives and almost no bias toward high lung function. Another major advantage with the whistle is that its estimation model is generalizable across devices and channels. In fact, it calculates PEF and FEV_1 directly, without any statistical modeling. For the patients with obstructive lung impairments such as asthma and COPD, the lung function measure that changes most drastically is FEV_1. If the patient only needs to track their FEV_1 with fine granularity (a common practice for many patients), SpiroCall can use a much simpler computation with a whistle, without any machine learning. Moreover, it will be easier to judge a valid effort because the shape of the curve is more faithfully represented.

SpiroCall's performance is promising as the mean performance loss due to use of the call-in service is only around 1%. The flexibility between channels and the possibility of using a whistle allows SpiroCall to make spirometry accessible. However, this only demonstrates the feasibility of sensing. It remains unclear how the user, in general, could use spirometers without any guidance from trained personnel. Although SpiroSmart tries to bridge this gap with a rich visual interface, it will be more difficult for SpiroCall to train the user. It is possible that in future work we could implement audio feedback between spirometry efforts, or have a health worker train the user before they are able to use SpiroCall independently.

CONCLUSION

In order to make spirometry more accessible, it is important to remove its dependence on smartphones. We introduced SpiroCall, a combination of call-in service and a simple whistle that turns *every* mobile phone in the world into a spirometer. The phone sends the audio data generated during a spirometry effort over the GSM voice channel and calculates the results on a central server. Our evaluation shows that we can use SpiroCall to reliably measure lung function in low resource regions. SpiroCall's call-in service's mean error is comparable to a clinical spirometer and does not degrade substantially when compared to local recordings made on a smartphone. The whistle helps in improving the performance with patients with degraded lung function. SpiroCall also serves as a demonstration that researchers can perform sensing on all mobile phones, not just smartphones, by leveraging the voice channel for data transfer.

REFERENCES

1. J M Bland and D G Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 8476: 307–10. Retrieved July 9, 2014 from http://www.ncbi.nlm.nih.gov/pubmed/2868172
2. Robert O Crapo, John L Hankinson, Charles Irvin, and Neil R. MacIntyre. 1994. Standardization of Spirometry. *American Journal of Respiratory and Critical Care Medicine*, 7.
3. Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2014. Least angle regression. *The Annals of Statistics* 32, 2: 407–499. Retrieved December 4, 2014 from http://projecteuclid.org/euclid.aos/1083178935

[1] www.spirometry360.org

4. Ericsson. 2015. *Ericsson Mobility Report: On the Pulse of the Networked Society*.

5. Lilian de Greef, Mayank Goel, Min Joon Seo, et al. 2014. BiliCam: Using Mobile Phones to Monitor Newborn Jaundice. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing 2014*, ACM Press. http://doi.org/10.1145/2638728.2638803

6. Maya R Gupta, Eric K Garcia, and Erika Chin. 2008. Adaptive local linear regression with application to printer color management. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 17, 6: 936–945. http://doi.org/10.1109/TIP.2008.922429

7. R J Knudson, R C Slatin, M D Lebowitz, and B Burrows. 1976. The maximal expiratory flow-volume curve. Normal standards, variability, and effects of age. *The American review of respiratory disease* 113, 5: 587–600. Retrieved December 4, 2014 from http://www.ncbi.nlm.nih.gov/pubmed/1267262

8. Jiri Kroutil, Alexandr Laposa, and Miroslav Husak. 2011. Respiration monitoring during sleeping. *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies - ISABEL '11*, ACM Press, 1–5. http://doi.org/10.1145/2093698.2093731

9. 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. *UbiComp'12*.

10. M R Miller, J Hankinson, V Brusasco, et al. 2005. Standardisation of spirometry. *The European respiratory journal* 26, 2: 319–38. http://doi.org/10.1183/09031936.05.00034805

11. Dennis R. Morgan and Michael G. Zierdt. 2009. Novel signal processing techniques for Doppler radar cardiopulmonary sensing. *Signal Processing* 89, 1: 45–66. http://doi.org/10.1016/j.sigpro.2008.07.008

12. Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless Sleep Apnea Detection on Smartphones. *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '15*: 45–57. http://doi.org/10.1145/2742647.2742674

13. Michael R Neuman. Vital signs: heart rate. *IEEE pulse* 1, 3: 51–5. http://doi.org/10.1109/MPUL.2010.939179

14. Sohail Rafiqi, Chatchai Wangwiwattana, Jasmine Kim, Ephrem Fernandez, Suku Nair, and Eric C Larson. 2015. Pupil Ware : Towards Pervasive C ognitive Load Measurement using Commodity Devices. *Proc. 8th Int. Conf. PErvasive Technol. Relat. to Assist. Environ.*

15. Alessandro Rubini, Andrea Parmagnani, and Michela Bondì. 2011. Daily variations in lung volume measurements in young healthy adults. *Biological Rhythm Research* 42, 3: 261–265. http://doi.org/10.1080/09291016.2010.505456

16. Hiroshi Sato and Kajiro Watanabe. 2000. Experimental study on the use of a vortex whistle as a flowmeter. *IEEE Transactions on Instrumentation and Measurement* 49, 1: 200–205. http://doi.org/10.1109/19.836334

17. Bernard Vonnegut. 1954. A Vortex Whistle. *The Journal of the Acoustical Society of America* 26, 1: 18–20.

18. Kajiro Watanabe and Hiroshi Sato. 1994. Vortex Whistle as a Flow Meter. *Proc. Advanced Technologies in Instrumentation and Measurement*, 1225–1228.

19. World Health Organization. 2015. Chronic obstructed pulmonary diseases (COPD). Retrieved September 9, 2015 from http://www.who.int/mediacentre/factsheets/fs315/en/

20. Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2: 301–320. http://doi.org/10.1111/j.1467-9868.2005.00503.x

21. 佐藤浩志, 大原昌幸, 渡辺嘉二郎, and 佐藤秀昭. 1999. Application of the Vortex Whistle to the Spirometer. 計測自動制御学会論文集 35, 7: 840–845.