Computerized Lung Sound Screening for Pediatric Auscultation in Noisy Field Environments

Dimitra Emmanouilidou, Eric D. McCollum, Daniel E. Park, and Mounya Elhilali

Abstract*—*Goal: **Chest auscultations offer a non-invasive and low-cost tool for monitoring lung disease. However, they present many shortcomings, including inter-listener variability, subjectivity, and vulnerability to noise and distortions. This work proposes a computer-aided approach to process lung signals acquired in the field under adverse noisy conditions, by improving the signal quality and offering automated identification of abnormal auscultations indicative of respiratory pathologies. *Methods:* The developed noise-suppression scheme eliminates ambient sounds, heart sounds, sensor artifacts, and crying contamination. The improved high-quality signal is then mapped onto a rich spectrotemporal feature space before being classified using a trained support-vector machine classifier. Individual signal frame decisions are then combined using an evaluation scheme, providing an overall patient-level decision for unseen patient records. *Results:* All methods are evaluated on a large dataset with $>$1000 children enrolled, 1–59 months old. The noise suppression scheme is shown to significantly improve signal quality, and the classification system achieves an accuracy of 86.7% in distinguishing normal from pathological sounds, far surpassing other state-of-the-art methods. *Conclusion:* Computerized lung sound processing can benefit from the enforcement of advanced noise suppression. A fairly short processing window size (<1 s) combined with detailed spectrotemporal features is recommended, in order to capture transient adventitious events without highlighting sharp noise occurrences. *Significance:* Unlike existing methodologies in the literature, the proposed work is not limited in scope or confined to laboratory settings: This work validates a practical method for fully automated chest sound processing applicable to realistic and noisy auscultation settings.**

***Index Terms*—Computerized lung sound interpretation, lung auscultation, multi-resolution analysis, noise suppression, noisy setting, pediatric.**

Manuscript received March 9, 2017; revised May 3, 2017; accepted June 8, 2017. Date of publication June 19, 2017; date of current version June 18, 2018. This work was supported in part by The National Institutes of Health under Grant R01HL133043 and in part by the Office of Naval Research under Grant N000141612045. The PERCH study was supported by Grant 48968 from The Bill & Melinda Gates Foundation to the International Vaccine Access Center, Department of International Health, Johns Hopkins Bloomberg School of Public Health. *(Corresponding author: Mounya Elhilali.)*

D. Emmanouilidou is with the Department of Electrical and Computer Engineering, Johns Hopkins University.

E. D. McCollum is with the Division of Pediatric Pulmonology, Johns Hopkins School of Medicine.

D. E. Park is with the International Vaccine Access Center, Johns Hopkins Bloomberg School of Public Health.

E. D. McCollum and D. E. Park are members of the Pneumonia Etiology Research for Child Health (PERCH) study team.

M. Elhilali is with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: mounya@jhu.edu).

Digital Object Identifier 10.1109/TBME.2017.2717280

I. Introduction

THE stethoscope is the most ubiquitous technology for accessing auscultation signals from the chest in order to evaluate and diagnose respiratory abnormalities or infections [1]. Since its invention in the early 1800s, the basic system has not changed much except for improvements in sound quality using shape modification and the introduction of enhanced materials. Despite its universal use, it remains an outdated tool, riddled with a number of issues. The stethoscope's value for clinical practice is limited by inter-listener variability and subjectivity in the interpretation of lung sounds. It is also restricted to well-controlled medical settings; the presence of background noise affects the quality of lung auscultations and may mask the presence of abnormalities in the perceived signal. It requires the interpretation of auscultation signals by properly trained medical personnel, which further limits its applicability within clinical settings without appropriate resources and medical expertise. These limitations are further compounded in impoverished settings and in pediatric populations. Close to 1 million children under five years of age die each year of acute lower respiratory tract infections (ALRI); more deaths than from HIV, malaria and tuberculosis combined [2]. Yet, access to medical expertise is not readily available and is further exacerbated by limited access to alternative diagnostic tools. Despite its limitations, the stethoscope remains a valuable tool in ALRI case management. Its potential is even more critical in resource-poor areas where low-cost exams are of paramount importance, access to complimentary clinical methods may be scarce or nonexistent, and medical expertise may be limited.

Computerized auscultation analyses (CAA) provide a reliable and objective assessment of lung sounds that can inform clinical decisions and may improve case management, especially in resource-poor settings. The challenges in developing such computerized auscultation analysis stem from two main hurdles. Firstly, there is great variability in the literature regarding a reliable description of lung signals and their pathological markers. For instance, adventitious sounds of *wheeze* have been reported

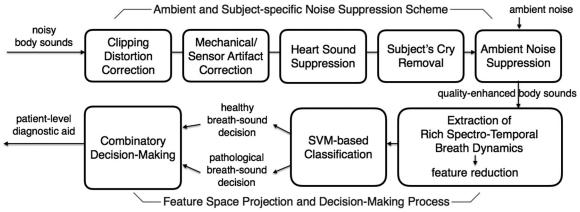

Fig. 1. Proposed integrated framework for complete auscultation solutions.

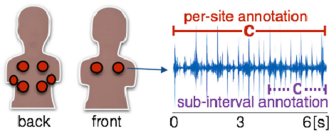

Fig. 2. Illustration of the 8 auscultation *sites* and the annotation process. A reviewer labeled the depicted *site* as crackles, C, in red/solid line, and then provided an indicative label of a crackling excerpt in purple/dashed line.

to span a wide range of frequencies varying within 100-2500 Hz or 400-1600 Hz; similarly *crackles* have been characterized as sounds with frequency content $<$2 kHz or $>$500 Hz or within 100-500 Hz [3], [4]. Secondly, ambient noise often contaminates the auscultation signal and masks important signature cues, as it often exhibits time-frequency patterns that greatly overlap with characteristic events in lung sounds [5].

Over the past few decades, few CAA approaches have been proposed to offer solutions to automated monitoring and diagnosis of lung pathologies. Nonetheless, the proposed approaches remain limited in their applicability, and tend to be confined to laboratory or well-controlled clinical settings or to simulated additive noise conditions [6]–[8]. These artificial settings greatly oversimplify environments in the field or the Emergency Department, where noisy and raucous clinical conditions incur unpredictable non-additive noise contamination. Few studies have explored analysis and classification techniques for breath sound diagnostics under more realistic clinical settings [9]–[13]; yet the majority suffers from limited patient evaluation or low protocol versatility. Unfortunately, the applicability of such methods to child auscultation is unknown and expected to be hampered by common pediatric challenges including irregular breathing, motion artifacts, crying or other body sounds that cannot be held back during examination. Finally, most proposed methods offer analysis techniques best suited to only identify context-specific pathological sound patterns [11]–[15].

A parallel challenge to the development of fully automated CAA systems is the need for hand-labeled information that can parse the respiratory phases in auscultation signals, identify specific signal instances with pathological markers as well as offer a reference medical interpretation of the auscultation signals. The need for such labeled ground-truth annotations is crucial for the development and training of supervised techniques, which explains why most studies are developed depending on it. Yet, a fully-annotated reference database is unrealistic because: (i) it is an extremely expensive and laborious effort in a large sample size; and (ii) it is not consistent with common medical practices where health care professionals rely on a global listening of the auscultation signal and recurrence of specific patterns indicative of pathologies while ignoring irrelevant information. Requiring an instant-by-instant labeling of hours of auscultation recordings is both unreasonable and impractical.

To tackle these challenges, we introduce an integrated scheme shown in Fig. 1 that (i) encompasses noise suppression to improve the signal quality, (ii) offers a rich feature representation to address the unpredictable nature of adventitious auscultation patterns, and (iii) provides patient-level assessment of pathological status by combining partial signal-level assessments without the need for exhaustively detailed annotations. For validation and evaluation, we use a large realistic dataset collected in developing countries in non-ideal rural and outpatient clinics. When it comes to distinguishing between normal vs. pathological lung sounds, we demonstrate the need for noise-free quality signals by using objective quality measures; we further demonstrate the advantages of the proposed feature extraction against state-of-the-art methods, which are shown here to lack the robustness to perform effectively on a diverse set of adventitious sounds, especially when noise events further mask the signal signatures.

Section II provides an overview of the digital data collection protocol and Section III presents the multi-step noise suppression scheme and evaluation. The rich feature space, classification and decision-making process follow in Section IV. Section V discusses patient diagnostic results as compared to other methods; and Section VI concludes the work with a discussion on the significance of these results.

II. Data Description and Preparation

All data and annotations were provided by the Pneumonia Etiology Research for Child Health (PERCH) study [16].

A. Data Collection

Digital auscultation recordings were acquired from children, ages 1 to 59 months (median age 7 ± 11.43 months), in outpatient or busy clinical settings in Africa (The Gambia, Kenya, South Africa, Zambia) and Asia (Bangladesh, Thailand). In total, 1157 children were enrolled into the digital auscultation study and were classified into one of the two categories: cases, having World Health Organization-defined severe or very severe pneumonia [17], or age-matched community controls, without clinical pneumonia.

The auscultation protocol called for recordings over 8 body locations (*sites*): four across the child's back, two in the axilla and two on the chest area (Fig. 2). To ensure two full breath cycles, at least 7 s of body sounds were obtained per *site*. A commercial digital stethoscope was used for data acquisition (ThinkLabs Inc. ds32a), sampling at 44.1 kHz. An independent Sony-ICD-UX71-81 microphone was affixed on the back of the stethoscope, recording concurrent ambient sounds. During examination the infant was seated, laid down or held to the most comfortable position.

B. Annotations

Nine expert reviewers (pediatricians or pediatric-experienced physicians) were enrolled for the annotation process. For each

TABLE I
AVAILABLE ANNOTATIONS OF PATIENTS' RECORDINGS

Annotation Label	Abnormal (Intervals with wheeze and/or crackles)	Normal (Intervals without wheeze nor crackles)
SUB-INTERVAL	annotated clip of arbitrary length found in abnormal *site* recordings of full or partial reviewer agreement	annotated clip of arbitrary length found in normal *site* recordings of full or partial reviewer agreement
PER-SITE (or SITE)	a *site* recording found abnormal by full or partial reviewer agreement	a *site* recording labeled normal by full or partial reviewer agreement
FULL-PATIENT	includes all *site* recordings of a patient if at least one *site* was found abnormal	includes all *site* recordings of a patient if all *sites* were found normal

patient recording, two distinct primary reviewers annotated the 8 sites (*per site* or *site* annotation) as being Normal or Abnormal (Table I), with an accompanying descriptor label: "definite", "probable" or "non-interpretable". A "definite" label was provided when the reviewer could interpret two or more full breaths with certainty. If only one breath could be interpreted with certainty or if two or more breaths could be interpreted with uncertainty, then a "probable" descriptor was given. If no full breath sounds could be distinguished (due to poor sound quality, technical errors, or unrecognizable contamination), a "non-interpretable" label descriptor was assigned.

The above process ensured that every *site* recording was assigned an annotation explaining breath sound findings, along with a confidence indicator for each finding. In case of disagreement between the two primary reviewers, more reviewers listened to the recording to resolve ambiguities, and provided additional labeling as needed (see [18] for details on the annotation process). Finally, within each *per site* label, reviewers were asked to specify a *sub-interval* label containing one segment of arbitrary length that best exemplified the given *per site* label (Fig. 2).

C. Datasets

Based on the *sub-interval* and *per site* labels, two types of data sets were created for the evaluation of this work:

1) *Sub-interval set:* including all patients' *sub-interval* recordings of arbitrary length, grouped into Normal and Abnormal (Table I, 1st row).
2) *Full patient set:* including all patients' records, grouped into Normal or Abnormal (Table I, 2nd-3rd row).

A few key-observations on the formed data groups: (i) adventitious events may still exist within a normal annotation, as long as their occurrence was not regarded a pathological lung sound; (ii) a *per site* recording was considered abnormal if there was full or partial agreement among reviewers over an abnormal annotation. Full or partial agreement means that a "definite" or "probable" presence of an abnormal sound was agreed by both primary reviewers or by at least two of the total reviewers. Augmenting the data sets to include both full and partial agreement cases ensured the minimization of excluded data, making the study more realistic, but at the expense of infusing uncertainty to the classification model; (iii) a patient record labeled as Abnormal (Table I, 3rd row), may contain one or more abnormal *sites* (Table I, 2nd row); (iv) patient records obtaining a "non-interpretable" label or failing to obtain full or partial agreement, were excluded from evaluation.

In total, 62 patients were excluded due to missing annotations, along with 29% of remaining *site* recordings, due to: "non-interpretable" labels, missing audio, recording malfunctions in one of the two microphones, or high disagreement among reviewer labels. The final included data set consisted of more than 250 hours of recorded lung sounds.

D. Preprocessing

All acquired recordings were low-pass filtered with an anti-aliasing 4th order Butterworth filter at 4 kHz cutoff; then resampled at 8 kHz and whitened to zero mean and unit variance. No crucial information loss was anticipated after down-sampling, given the nature of the recorded signals and the suggested guidelines [19]: normal respiratory sounds are typically found between 50–2500 Hz, tracheal sounds can reach energy contents up to 4000 Hz, abnormal sounds including wheeze, crackles, stridors, squawks, rhonchi or cough exhibit a frequency profile below 4000 Hz, and heart beat sounds can be found in the range of 20–150 Hz.

III. SIGNAL ENHANCEMENT

Auscultation recordings acquired in busy clinical settings are often prone to environmental noise contamination, and result in inherent difficulties for both the physician and computerized methods. PERCH recordings were also heavily corrupted by contamination of various noise sources such as family members talking close to the patient, children crying in the waiting room, musical toys, vehicle sirens, mobile or other electronic interference, and other. An effective noise suppression scheme was developed below, crucial for suppressing exterior contamination before further analysis.

A. Clipping Distortions

Clipping distortions are produced when the allowed amplitude range of the stethoscope sensor or recording device is exceeded. The incoming sound signal is then truncated, enforcing the loss of high amplitude content and resulting in significant distortion. Both the time and spectral signal signatures are heavily affected by the non-trivial high frequency harmonics formed. Clipped regions were identified as consecutive time samples with constant maximum-value amplitude, up to a small 3% perturbation tolerance [Fig. 3(a)]. Then, the identified regions were repaired using spline piece-wise cubic interpolation; given the brief duration of clipping intervals (a few consecutive data samples), this method was adequate for replacing the distorted portions without distorting the physiological sound signal.

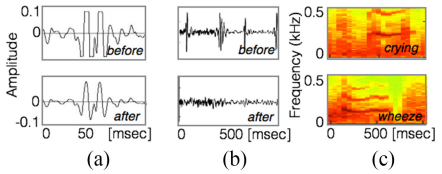

Fig. 3. (a) Waveform of a lung sound excerpt distorted by clipping (flat amplitude regions in panel "before"), and the corresponding output of the correction algorithm (panel "after"); (b) waveform of a lung sound excerpt illustrating the effects of the heart sound interference suppression; notice the suppressed heart sound patterns (panel "after") when compared to the original waveform ("before"); (c) two spectrogram representations of lung sound excerpts illustrating the inherent difficulty in differentiating between wheezing patterns and crying contamination.

B. Mechanical or Sensor Artifacts

Mechanical or sensor noise is usually generated when the physician moves the stethoscope to various body locations or when the stethoscope is unintentionally and abruptly displaced. This is a common distortion, and especially prominent during pediatric auscultation. Sharp stethoscope movements are typically associated with skin friction and produce irregular short-time broadband energy bursts in the sound signal, resembling profiles of abnormal lung sounds such as crackles. In the current dataset, the stethoscope transition noise was identified as follows: the auditory spectrogram (ASP) representation was calculated on an 8 ms window (described in details later in (4)), and normalized to [0, 1]. Mostly interested in broadband events, the region of interest ROI_{ASP} within the ASP spectrum, was defined as high spectral content above 1 kHz, with a span greater than 1.5 kHz. Consecutive frames, of 8 up to 100 ms, exhibiting high energy content within ROI_{ASP} were identified and discarded.

C. Heart Sound Interference

In the context of auscultation recordings, heart sounds (HS) are yet another added component masking respiratory sounds. Heart signal suppression has been addressed in several studies using various techniques including wavelets and Short Time Fourier Analysis [20], [21]. In order to maintain the integrity of the lung sounds, particularly any adventitious events, a conservative approach was used here, utilizing a wavelet multi-scale decomposition [22].

(i) HS identification: The original lung sound signal was band-pass filtered in [50, 250] Hz and down-sampled to 1 kHz, using a 4th order Butterworth filter. This step enhanced heart beat components by suppressing lung sounds and noise components outside this range. Next, the discrete Static Wavelet Transform (SWT) was obtained at depth 3, using Symlet decomposition filters (due to their appropriate shape): after Detail $D_j(t)$, and Approximation $A_j(t)$ coefficients were obtained, signals did not undergo down-sampling, which allows for the time-invariance of the transform. Signal reconstruction was then easily obtained by averaging the inverse wavelet transforms [23]. Let $SWT_j\{s(t)\}$ be the wavelet decomposition at the jth scale level of the lung sound signal $s(t)$ and $A_j(t)$ be the obtained normalized approximation coefficient. Then $P_{1:J}(t)$ is the multiscale product of all J approximation coefficients, defined in (1). Intervals achieving high values for $P_{i:j}$, were identified as heart sounds and were replaced using an ARMA model.

$$P_{i:j}(t) = \prod_{j=1}^{J} A_j(t)/\max(|A_j(t)|) \tag{1}$$

(ii) HS replacement: Assuming that lung sounds are locally stationary, an ARMA model was employed to replace missing data of $x(n)$ using past or future values. First a stationarity check - explained next - was performed on the neighboring area of the removed segment. If the post-neighboring segment was found non stationary, then a forward linear prediction model was used (2a); otherwise, a backward model was used (2b):

$$\hat{x}(n) = -\sum_{k=1}^{p} \alpha_p(k) x(n-k) \tag{2a}$$

$$\hat{x}(n-p) = -\sum_{k=0}^{p} \beta_p(k) x(n-k) \tag{2b}$$

where $\{-\alpha_p(k), -\beta_p(k)\}$ denote the prediction coefficients of the order-p predictors. Solving for the coefficients by minimizing the mean-square value of the prediction error $\{x(n) - \hat{x}(n)\}$ leads to the normal equations involving the autocorrelation function, $\gamma_{xx}(l)$: $\sum_{k=0}^{p} \alpha_p(k)\gamma_{xx}(l-k) = 0$, with lags $l = 1, 2, .., p$ and coefficient $a_p(0) = 1$. The Levinson-Durbin algorithm was used to efficiently solve the normal equations for the prediction coefficients. The order of each linear prediction model was determined by the length of the particular heart sound gap, using an upper bound of $p_{\max} = 125$ ms.

For the stationarity check, the two neighboring intervals around the missing data, of length $T_i = 200$ ms, were partitioned into M non-overlapping windows of length L. Using the Wiener-Khintchine theorem, the power spectral density of the m-th segment, $\Gamma^m_{xx}(l)$, was computed via the multitaper periodogram and the following spectral variation measure was introduced [24]

$$V(x) = \frac{1}{ML} \sum_{l=0}^{L-1} \sum_{m=0}^{M-1} (\Gamma^m_{xx}(l) - \frac{1}{M} \sum_{k=0}^{M-1} \Gamma^k_{xx}(l))^2 \tag{3}$$

with $V(x) = 0$ signifying a wide-sense stationary process.

Among identified HS intervals, only the very prominent ones were chosen to be replaced, i.e. the ones achieving increased product values $P_{i:j} > 0.2$. Additionally, if the peak-to-peak interval for identified heart sounds was too short for pediatric standards (<0.28 s), then the corresponding identified regions (possibly indicative of other adventitious sounds) were not replaced. Fig. 3(b) shows an example of a heart sound suppressed segment.

D. Subject's Intense Crying

Depending on the cause of irritation, infants and young children can broadcast crying vocalizations of varying temporal and frequency signature modes [25], [26]: phonation, consisting of the common cry with a harmonic structure and a fundamental

frequency ranging in 350-750 Hz; hyperphonation, a sign of major distress or pain, also harmonically structured but with rapidly changing resonance and a shifted fundamental frequency of 1–2 kHz or higher; and dysphonation (beyond the scope of this work), a sign of poor control of the respiratory cycle, containing aperiodic vibrations.

Because of their spectral span and harmonic structure, instances of phonation and hyperphonation cry were identified using properties of the signal's time-frequency representation. However, since adventitious lung sounds (particularly wheezes) can produce patterns of similar or overlapping specifications [Fig. 3(c)], here the focus was on longer, intense crying intervals bearing limited value for clinical assessment.

For the detection of phonation mode cry: (i) The ASP representation was calculated for every 8 ms frame (described in details later in (4)). A pitch estimate for every frame was calculated, using an adaptation of a template matching approach [27]. Each spectrogram slice was compared to an array of pitch spectral templates, generated by harmonically-related sinusoids, modulated by a Gaussian envelope. The dominant pitch per frame was then extracted and the average pitch (excluding 20% of distribution tails) constituted the resulting pitch estimation per region. Frames with an extracted pitch lower than 250 Hz were immediately rejected. To avoid confusion with possible adventitious occurrences during inspiration or expiration, an identified interval was required to be of duration $T_{\rm dur} > 600$ ms, considering respiratory rate standards for infants [28]; typical rates in the current dataset were 18–60 breaths per minute. (ii) Features of spectro-temporal dynamics (6)–(8) were extracted from all candidate time-segments, and fed to a pre-trained, binary SVM classifier using radial basis functions, to distinguish crying from other voiced adventitious sounds like wheezes.

For hyperphonation, simpler steps were required as lung sounds were unlikely to overlap with this type of cry: regions with high ASP spectral content above 1 kHz, and exceeding a duration of $T_{\rm dur}$, were detected as hyperphonation cry.

In total, 20% of all recorded lung signals were identified as phonation or hyperphonation cry, demonstrating the necessity of such processing step.

E. Ambient Noise

Lung auscultation is highly vulnerable to ambient noise interference, especially when patients are examined in busy clinics or non-soundproof rooms. Natural occurring environmental sounds, vehicle sounds, electronic machinery sounds, phones ringing, conversational speech or distant crying all fall under the umbrella of ambient noise commonly found in realistic auscultation protocols, like the PERCH study.

A modified spectral subtraction scheme was employed for suppressing such complex noise contamination. The general spectral subtraction scheme assumes a known measured signal quantity s (noisy lung sounds) to be comprised of two signal components $s = x + d$: the unknown desired signal x (pure clean lung sounds) and a known or approximated interference signal d (ambient sound pick-up signal). The algorithm operates in the spectral domain, in short frames to allow for short-term

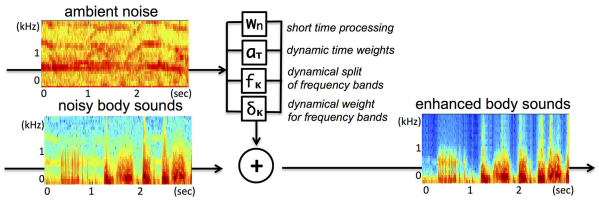

Fig. 4. Pipeline illustration of the ambient noise suppression scheme.

stationarity assumptions, and the content of the clean signal is obtained by $|S|^2 = |X|^2 - |D|^2$, where X, S, D correspond to the short time discrete Fourier Transform (STFT) of x, s, d respectively.

An extension of this general framework to chest sounds would not be readily sufficient or effective, due to the intricate nature of these signals. The design above was extended as part of our previous work [9], to account for (i) the preservation of the sensitive lung sound content present in both low and high frequencies (ii) localized frequency treatment, by adaptively splitting the frequency range and ensuring robustness over unpredicted noise environments; (iii) localized time window treatment, by using the local Signal To Noise Ratio (SNR) information to individually adjust the amount of subtracted information; this way, both slow and fast-varying contamination can be treated; and finally account for (iv) the elimination of reconstruction distortions such as "wind tunnel" noise effects, by smoothing signal estimates along adjacent frames and frequency bands. This modified, adaptive spectral-subtraction scheme was validated by 17 medical experts, who confirmed that the valuable breath sound was faithfully preserved in the recovered signals, while the ambient noise was successfully suppressed (Fig. 4).

F. Objective Quality Assessment of Enhanced Lung Sounds

A subjective sound quality assessment before and after the ambient noise suppression scheme had been previously shown, by enrolling medical experts that evaluated sounds based on their quality and preservation of the lung sound content [9]. Here we attempt a sound quality assessment offered by the overall noise suppression scheme, based on objective measures. The choice of an appropriate metric is not a trivial task since (i) there is no available standardized method for evaluating quality of lung sound content (ii) most quality measures proposed for speech or sound enhancement require knowledge of the true clean signal [29], [30], which in our case, would be the true clean lung sound of the individual patient, a quantity that is unknown for non-simulated environments.

In absence of the true underlying lung sound content, here we assess each step of the proposed noise-suppression framework by comparing the amount of shared information with the picked-up background noise. Evidently, this approach is not a conventional measure for signal quality improvement, but offers a practical alternative to quality assessment adjusted to the problem at hand. It assesses how much information is shared between the background or subject-specific noise and the

signals before, during and after the sound enhancement process. Two objective metrics were explored:

1) *Normalized-Covariance*:

$$NCM = \sum_{k=1}^{K} w_k SNR^N(k) \Big/ \sum_{k=1}^{K} w_k$$

NCM is a measure used specifically for estimating speech intelligibility (SI) by accounting for audibility of the signal at various frequency bands. It is a measure based on the speech-based Speech Transmission Index (STI). It captures a weighted average of a Signal to Noise quantity SNR^N, calculated from the covariance of the envelopes of the two signals over different frequency bands k [31] and normalized to [0, 1]. A value equal to 1 is achieved when the signals under comparison are identical. The band-importance weights w_k followed ANSI-1997 standards [32]. Though this metric is speech-centric, it is constructed to account for audibility characteristics of the human ear hence reflecting a general account of improved quality of a signal as perceived by a human listener.

2) *Three-Level Coherence Speech Intelligibility Index*:

$$CSII_x = \frac{1}{T}\sum_{\tau=1}^{T}\left\{\sum_{k=1}^{K} w_k SNR^N_{ESI}(k,\tau) \Big/ \sum_{k=1}^{K} w_k\right\}$$

The CSII metric is also a speech intelligibility-based metric, based on the ANSI standard for the Speech Intelligibility Index (SII). Unlike NCM, CSII uses the signal-to-residual SNR^N_{ESI}, an estimate of Signal-to-Noise ratio in the spectral domain, for each frame $\tau = 1, ..., T$; it is calculated using the ro-ex filters and the Magnitude-Squared Coherence (MSC) followed by [0, 1] normalization, with a value of 1 signifying identical signals. A 30 ms Hanning window was used and the three-level CSII approach divided the signal into low, mid, and high-amplitude regions, using each frame's root mean square (rms) level information. The high-level region $CSII_{\text{high}}$ consisted of segments at or above the overall rms level of the whole utterance. The mid-level $CSII_{\text{mid}}$ consisted of segments ranging from the overall rms level to 10 dB below, and the low-level $CSII_{\text{low}}$ consisted of segments ranging from rms -10 dB to rms -30 dB [33].

IV. Classification Model

A. Acoustic Analysis

After signal enhancement, an analysis of the joint spectral and temporal characteristics of the auscultation signal was performed. A biomimetic approach was employed, and the acoustic signal was projected onto a high-dimensional space spanning time, frequency, as well temporal dynamics and spectral modulations. The analysis followed the model proposed in [34], [35] by adapting it to auscultation signals; and is summarized below:

The auscultation signal $s(t)$ was first analyzed through a bank of 128 cochlear filters $h(t; f)$, with 24 channels per octave. These filters were modeled as constant-Q asymmetric band-pass filters and tonotopically arranged with their central frequencies logarithmically spaced. Then, signals were pre-emphasized by a temporal derivative and spectrally sharpened using a first-order difference between adjacent frequency channels, followed by half-way rectification and a short-time integration $\mu(t;\tau)$, with $\tau = 8$ ms. The result was an enhanced representation, the auditory spectrogram:

$$y(t,f) = \max(\partial_f \partial_t\ s(t) *_f h(t,f), 0) *_t \mu(t;\tau) \tag{4}$$

This time-frequency representation was further expanded to extract signal modulations using a multiscale wavelet analysis, akin of processes that take place in the central auditory pathway, particularly at the level of auditory cortex [35]. This analysis yields a rich feature representation that captures intrinsic dependencies and dynamics in the lung sound signals along both time and frequency. This stage is implemented by filtering the auditory spectrogram $y(t, f)$ through a bank of modulation-tuned filters G, selective to specific ranges of modulation in time (rates $\mathfrak{r}$ in Hz) and in frequency (scales $\mathfrak{s}$ in cycles/octave or c/o):

$$G_+(t,f;\mathfrak{r},\mathfrak{s}) = A^*(h_r(t;\mathfrak{r}))A(h_s(f;\mathfrak{s})) \tag{5a}$$

$$G_-(t,f;\mathfrak{r},\mathfrak{s}) = A\ (h_r(t;\mathfrak{r}))A(h_s(f;\mathfrak{s})) \tag{5b}$$

where A(.) indicates the analytic function, $(.)*$ is the complex conjugate, and $+/-$ indicates upward or downward orientation selectivity in time-frequency space, i.e., detecting upward or downward frequencies sweeping over time: a positive rate corresponds to downward moving energy contents and a negative rate corresponds to upward moving energy contents. The seed functions $h_r(t)$ and $h_s(f)$ were shaped as Gamma and Gabor functions respectively

$$h_r(t) = t^3 e^{-4t} cos(2\pi t), \quad h_s(f) = f^2 e^{1-f^2} \tag{6}$$

A filter bank was constructed by dilating the seed function and creating 31 filters of the form $h_r(t;\mathfrak{r}) = \mathfrak{r} h_r(\mathfrak{r}t)$ to capture slow/ fast temporal variations for modulations $\mathfrak{r} = 2^{[1.4:0.22:8]}$; and 21 filters of the form $h_s(f;\mathfrak{s}) = \mathfrak{s} h_s(\mathfrak{s}f)$, to capture narrow/broadband spectral content, with $\mathfrak{s} = 2^{[-5:0.4:3]}$. Each modulation filter output modeled the response of differently-tuned filters, mapping the time waveform onto a high-dimensional space:

$$r_\pm(t,f;\mathfrak{r},\mathfrak{s}) = y(t,f)\ *_{t,f}\ G_\pm(t,f;\mathfrak{r},\mathfrak{s}) \tag{7}$$

where $*_{t,f}$ corresponds to convolution in time and frequency and $G_\pm$ is the 2D modulation filter response. The final representation was obtained by integrating the response along time, achieving a frequency-rate-scale description:

$$R_\pm(f;\mathfrak{r},\mathfrak{s}) = \int_t r_\pm(t,f;\mathfrak{r},\mathfrak{s})\ \delta t \tag{8}$$

Note that even though the time axis is integrated in the equation above, details of the temporal changes in the signal are captured along the rate axis $\mathfrak{r}$.

B. Reduction of Feature Space Dimension

To reduce the size of the feature space, tensor Singular Value Decomposition (SVD) was used. Data was unfolded along each dimension of the SVD space, created by the training data set

only. Let R be the feature tensor of order 3 seen above, where the R_- axis is concatenated with the R_+ axis, so that $R \in \mathbb{R}^{d_1 \text{x} d_2 \text{x} d_3}$, where $d_1 = 128$ for the frequency axis, $d_2 = 31 \times 2 = 62$ for both $\pm$ rates, and $d_3 = 21$ for scales. When unfolding R along mode (dimension) 1, an order-2 tensor (or matrix) was created, $R^{(1)}$, of dimensions d_1x(d_2xd_3). Similar order-2 tensors were also created when unfolding along dimension 2 and 3, creating matrices $R^{(2)}$ and $R^{(3)}$. Singular value decompositions were obtained for each of the mode unfoldings $R^{(n)}$, for $n = 1, .., 3$ as:

$$R^{(n)} = U^{(n)} \Sigma^{(n)} V^{(n)\,T}$$

For mode-1 unfolding, $\Sigma^{(1)}$ is a diagonal matrix of dimension r, with the nonzero singular values on its diagonal; $r \leq \min\{d_1, (d_2 \text{x} d_3)\}$ is the rank of $R^{(1)}$, i.e. the dimension of the space spanned by the columns or rows of $R^{(1)}$ and $U^{(1)}$ and $V^{(1)\,T}$ are unitary matrices. The singular values in $\Sigma^{(1)}$ are presented ranked, as $\sigma_1^{(1)} > \sigma_2^{(1)} > ... > \sigma_r^{(1)} > 0$. Similar expressions were obtained for mode-2 and mode-3 decomposition. For each $R^{(n)}$, only components capturing up to 99% of the total variance were kept (i.e. $r^{(n)} = \arg\min_x \; f(x) := \{\sum_{i=1}^{x} \sigma_i^{(n)} \geq 0.99 \; | x = 1, ..., d_n\}$. The final space projection was achieved by tensor-matrix multiplication (mode-n product), significantly reducing the feature dimensions from 128 $\times$ 62 $\times$ 21 to about 5 $\times$ 3 $\times$ 3 (exact dimension may vary depending on the training subset).

C. Auscultation Classification

The classification of feature vectors into Normal vs. Abnormal was obtained using a soft-margin non-Linear Support Vector Machine (SVM) classifier. Let $\mathbf{x}$ be the matrix comprising of all x_i SVD-projected feature vectors $\in \mathbb{R}^r$, where $r = \prod_{n=1}^{3} r^{(n)}$; and let Φ be a kernel mapping where data is believed to be separable, so that $\Phi(\mathbf{x}) : \mathbf{x} \rightarrow \Phi(\mathbf{x})$, mapping data from $\mathbb{R}^r \rightarrow \mathbb{R}^D$, $D > r$. Given knowledge of data points $\mathbf{x}$, and their true class y, a binary SVM classifier, seeks to learn an optimal hyperplane $\mathbf{w}^T \; \Phi(\mathbf{x})$, $\mathbf{w} \in \mathbb{R}^D$, where

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b$$

is the output class participation ($f(\text{x}_i) = \pm 1$) of example x_i; $b = +1 - \mathbf{w}^T \Phi(\mathbf{x})$ for examples in class 1; $b = -1 - \mathbf{w}^T \Phi(\mathbf{x})$ for examples in class -1; and $|\mathbf{w}| = 1$. The optimal hyperplane is found by solving the unconstrained quadratic minimization problem over $\mathbf{w}$:

$$\min_{\mathbf{w} \in \mathbb{R}^D} \; ||\mathbf{w}||^2 + C \Sigma_i^N \max\left(0, 1 - y_i \; f(x_i)\right)$$

where N is the number of learning data points and C is a regularization parameter. The second term represents the loss function, where $y_i f(x_i) > 1$ if a data point x_i falls over the correct side of the separating hyperplane margin and $y_i f(x_i) = 1$ if it falls on the margin; finally, $y_i f(x_i) < 1$ if the data point falls on the wrong side of the margin. The optimization problem can also be expressed in its dual form:

$$f(\mathbf{x}) = \Sigma_i^N \alpha_i y_i K(x_i, x) + b$$

$$\max_{a_i \geq 0} \Sigma_i a_i - \frac{1}{2} \Sigma_{j,k} a_j a_k y_j y_k k(x_j, x_k)$$

subject to $0 \leq a_i \leq C, \forall i$, and $\sum_i a_i y_i = 0$. In the present work, radial-basis kernels (RBF) were used $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) = \exp(-|x(i) - x(j)|^2)$. This way, only the learning of *N*-dimensional vector **a** is needed, avoiding the learning of *D*-dimensional $\mathbf{w}$ in the primal problem.

D. Timescale of Diagnosis

Choosing the timescale (analysis window) over which to perform classification is a nontrivial task. An ideal parsing of the signal would require a window segmentation aligned to the breathing cycle. While this is often the chosen parsing method in studies of limited data [7], [36], [37], it is an impractical solution for large datasets recorded in the field: obtaining pre-annotated breath cycles for all subjects is unrealistic and cannot be automated in a straight-forward manner, especially when considering the irregularity of infant breathing. Alternatively, one could opt for a fixed-size window, which will likely have an impact on the classification outcome. On one end of the spectrum, a very short window will highlight short adventitious events, at the expense of great heterogeneity among training data, especially under noisy conditions. On the other end of the spectrum, a very long window would capture average characteristics of normal vs. abnormal lung sound events but could blend details pertaining to short pathological patterns. We investigated a variety of analysis windows ranging from shorter to longer duration: $W_i \in$ [0.3, ..., 5] s with 50% overlap.

E. Evaluation of Classification Results

A closely related issue is the timescale of *evaluating* classification results. The available auscultation dataset contained one annotation per each 7 s recording *site* (see Section II-C); full-scale, extensive annotations of all sounds of interest were not available and are not a realistic feature, thus, we propose the following algorithmic performance evaluation technique:

1) Sub-Interval Evaluation: (used for study comparison in Section V.C): all arbitrary-length *sub-interval annotations* of all available patient records were included in this dataset, grouped into two groups (Normal/Abnormal). A decision for each sub-interval clip was made by the SVM classifier, leading to performance evaluation on the *sub-interval* level;

2) Full Patient Evaluation: (used for extended evaluation of proposed method in Section V.B): this dataset combined individual frame decisions of each *site* into an overall patient decision. This is not a trivial task, and our approach was designed to be highly sensitive to abnormal occurrences. First, all grouped *site* recordings were split into individual frames of length $W_i \in [0.3, ..., 5]$ s with 50% overlap, and a classifier decision was made at the frame level. Next, a combined decision for each *site* was obtained as follows: a *site* received an abnormal output label if at least (i) 2 consecutive intervals of α

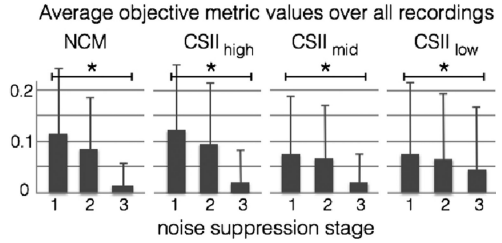

Fig. 5. Objective quality metrics illustrating the amount of shared information between the ambient noise and the different noise suppression stages. Low values indicate that signals under comparison have less content in common. Standard deviation error bars show variation among all *site* recordings. The asterisk (*) indicates that the trends across all stages of denoising are statsitically significant at the 0.0005 level, using both ANOVA and kruskal-wallis tests.

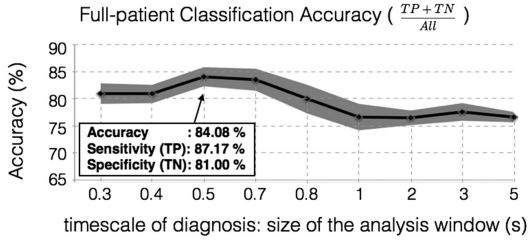

Fig. 6. Final patient-classification results. Performance was calculated based on the *full-patient decision*; Accuracy = (TP+TN)/All %, where TP: number of True Positives (abnormal patients), TN: number of True Negatives (normal patients), All: total number of patients. Grey shading depicts the standard deviation in patient accuracy among 10 MC runs.

duration were found to be abnormal by the classifier or if at least (ii) $\beta\%$ of all overlapping frames were found to be abnormal; (this approach was partially inspired by the annotation protocol that the medical experts followed - Section II-B). Finally, a full patient record was assigned an abnormal label if at least one of its *sites* was found to be abnormal; otherwise the patient record was assigned a normal output label. For each time window W_i, parameters α and β were optimized in [0, 2] s and [30, 70]% respectively.

V. Results

A. Objective Quality Assessment of Enhanced Lung Sounds

Objective metrics NCM and CSII were employed to quantify improvements to the signal quality before, during and after the signal enhancement. The metrics were calculated between the clipping corrected ambient noise signal and (i) the original clipping corrected noisy lung sound (Stage 1 in Fig. 5); (ii) the processed lung sound after additionally applying sensor artifact correction, heart sound suppression and crying elimination (Stage 2 in Fig. 5); and (iii) the fully enhanced lung sound after applying all noise suppression steps including the ambient sound suppression (Stage 3 in Fig. 5). All metrics demonstrated an attenuating trend in the amount of information shared between ambient noise and processed signals, along various stages of the noise suppression scheme. An analysis of statistical signficance of these trends indicate that they are significant at the 0.0005 level for both ANOVA and kruskal-wallis tests. The attenuating trend is an indication that the processed lung signal shares less content with the noise, when compared to the original lung recording. It further depicts the necessity for efficient noise suppression techniques which can play an important role in improving the quality of auscultation signals and facilitating the work of physicians for diagnostic purposes, allowing data re-usability for educational or training purposes and also improving further computerized analysis with the extraction of more robust features.

B. Full Patient Diagnostics

After combining the noise suppression scheme with the rich feature analysis and decision integration, the accuracy of the complete system was assessed for patient-level decisions, using the full-patient evaluation process of Section IV-E2. As outlined earlier, the system performance depends crucially on the choice of analysis window W_i (timescale of diagnosis). Fig. 6 shows the system accuracy for different analysis windows. On one hand, large windows >1 s capture the coarse characteristics of the lung sounds at the expense of the refined detection of adventitious events such as crackle which can be very localized in time and are integrated in these longer time windows. Such coarse analysis yields an accuracy of about 77%. On the other hand, a very short analysis window <0.5 s can be sensitive to very small or transient changes in the signal hence failing to track sustained patterns of interest such as wheezes which tend to be very musical in nature and can last few hundreds of milliseconds. Such short windows also yield a smaller drop in accuracy. Overall, it is observed that a balanced time window of about 0.5 s is preferred as it balances the detailed analysis with the tracking of events of interest. Using the recommended 0.5 s, our proposed integrated system yields an overall patient-level accuracy of 84.08% in Fig. 6. The shaded area shows the standard deviation in accuracy over 10 Monte-Carlo runs.

C. Comparison With Other Methods

The effectiveness of the proposed biomimetic features was furthered explored via a comparison with state of the art methods in the literature. Palaniappan *et al.* demonstrated the use of the Mel-frequency cepstral coefficients (MFCCs) for capturing spectral characteristics of normal and pathological respiratory sounds [38]. MFCCs are powerful features commonly used in audio signal processing, particularly in speech applications; it is a type of nonlinear cepstral representation calculated on a mel frequency axis, which approximates spectral perception of human listeners [39]: first, the logarithm of the Fourier transform was calculated using the mel scale followed by a cosine transform. One MFCC coefficient was obtained per frequency band, and in total, 13 MFCCs were derived for each data excerpt, averaged over a processing window of 50 ms with 25% overlap. This method is referred to as $MFCC\_P$.

In a different study by Jing *et al.* [40], a new set of discriminating features was proposed for identifying adventitious events in respiratory sounds, based on spectral and temporal signal characteristics. The features were extracted from a refined spectro-temporal representation, the Gabor time-frequency (Gabor TF)

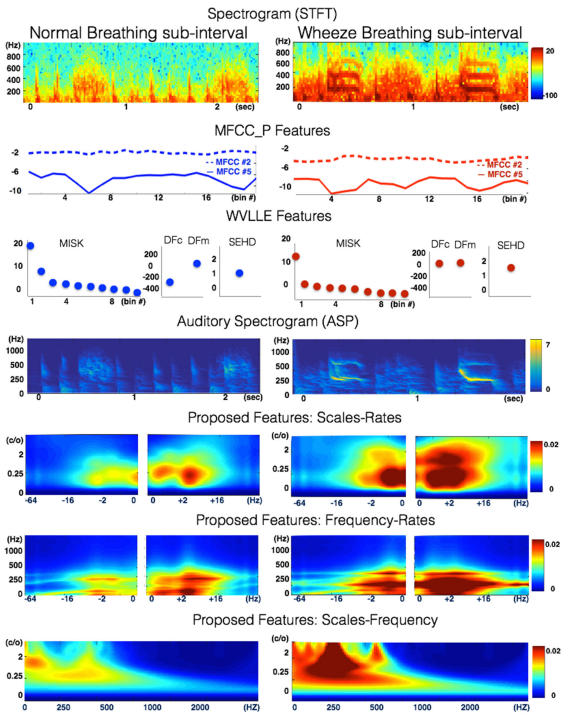

Fig. 7. Comparison of feature extraction methods for a normal (left) ad a wheeze (right) lung sound. Row 1: time-frequency breath characteristics; Row 2: binned MFCC coefficient #2 (75 Hz) and #5 (200 Hz) extracted as part of the *MFCC*$_P$ method. Row 3: features MISK, DFc, DFm and SEHD, extracted as part of the $WVILLE$ method; Rows 4–7: the proposed discriminating features including the auditory spectrogram ASP and the combined spectral and temporal breath dynamics. Notice the high discriminatory nature of the proposed features: the wheezing breath is highlighted with high energy concentration in the Scales-Rates plot $\sim$1 c/o, capturing its harmonic structure, and in the Frequency-Rates and Scales-Frequency plots $\sim$200 Hz, capturing its pitch. Comparatively, the normal breath exhibits much lower temporal and spectral dynamics.

distribution. As the order of the Gabor TF representation increases, it converges to a Wigner-Ville distribution, and we used the latter to extract multiple features from each frequency band, as proposed by the authors: MISK: mean instantaneous kurtosis, used as feature for adventitious sound localization; DFc and DFm denoting the contrast and minimum value of the calculated discriminating function, used for signal predictability features; and SEHD: sample energy histogram distortion, used as a nonlinear separability criterion for breath discrimination. This method is referred to as $WVILLE$.

For a comparison focused on the effectiveness of the extracted features, we used the data pool created from the *sub-interval* annotations (Section IV-E1) of all subjects in the PERCH database, after full signal enhancement. Recall that the *sub-interval* annotations can be of arbitrary length (with an average duration of 1.8 s in this database). In order to create a relatively uniform database, the intervals were clipped or augmented to 2 s, while intervals shorter than 1 s were discarded.

Fig. 7 illustrates the differences of all the feature extraction techniques, as applied on a normal and a wheezing lung sound clip. Row 1 depicts the sound spectrograms calculated on a 30 ms, 50% overlap window simply shown here for reference.

TABLE II
COMPARATIVE CLASSIFICATION RESULTS

	Sensitivity (TP)%	Specificity (TN)%	Accuracy%
PROPOSED	86.82 ($\pm$0.42)	86.55 ($\pm$0.36)	86.67
MFCC_P	91.88 ($\pm$0.36)	53.40 ($\pm$0.74)	72.64
WVILLE	63.86 ($\pm$0.55)	58.47 ($\pm$0.60)	61.16

*Performance based on *sub-interval* decision.

Row 2 shows MFCC coefficients #2 and #5 tuned at 75 Hz and 200 Hz respectively, extracted by MFCC$_P$ method. Row 3 shows the WVILLE features: the 10 maximum average instantaneous kurtosis values (MISK); the minimum achieved value of the enclosed discriminating function (DFm) and its center-surround contrast (DFc); and the histogram distortion value (SEHD). Row 4 shows the ASP spectrogram used in the proposed method for extracting the spectro-temporal breath dynamics. Rows 5-7 depict the 3-dimensional Frequency-Rate-Scale space, shown on individual two-dimensional projections. Notice the high discriminatory nature of the proposed set of features: the wheezing breath is highlighted by the presence of strong energy components $\sim$1 c/o in the Scales-Rates plot (capturing its harmonic structure), and the energy concentration around 200 Hz along the y-axis of the Frequency-Rates and Scales-Frequency space (capturing its pitch). Compared to the normal breath, the wheezing breath exhibits much higher temporal dynamics as captured by the rates axis.

The RBF SVM classifier was used for all compared methods evaluated on a 10-fold cross validation and 20 Monte Carlo repetitions. Subjects in the training and testing sets were again, mutually exclusive, to avoid classification bias. Recall, that while a normal annotation rules out wheeze or crackle occurrences, the lack of other abnormal sounds such as upper respiratory sounds (URS) or remaining noise cannot be guaranteed, adding real life challenges to the data. Comparative results are shown in Table II, with the accuracy index depicting the average of sensitivity (True Positives Rate) and specificity (True Negatives Rate). The superiority of the proposed feature extraction method was revealed; the rich spectro-temporal space spans intricate details in the lung signal and results in better discriminatory features. Importantly, the proposed features appear to be equally robust in identifying normal and abnormal breath sounds without any bias. In contrast, low accuracy percentages of the $WVILLE$ method are noticeable; the $WVILLE$ features were designed to detect unexpected abnormal patterns within specific breath context, and the feature space seems to lack the ability of separating respiratory-related abnormal sounds from noise-related sounds, signal corruption, or breaths containing possible URS. $MFCC_P$ features were better qualified for identifying abnormal breaths, but when it came to normal segments, both $WVILLE$ and $MFCC_P$ fail to distinguish from noise or other contamination. The $MFCC_P$ and $WVILLE$ methods were previously reported in [38] and [40] to obtain an average accuracy of 77.42% and Area Under the Curve accuracy of 95.60% respectively, in distinguishing normal from pathological lung sounds. However findings of the current work clearly illustrate

the inherent difficulty of these feature extraction methods to generalize findings to more realistic or challenging databases and auscultation scenarios.

VI. Conclusion

Over the last decades, there has been an increased interest in computer-aided lung sound analysis. Despite the enthusiasm about possibilities in automated diagnosis, the literature is still shy in tackling real-life challenges. The presented method addresses some of these limitations by proposing a robust discriminative methodology for distinguishing normal and abnormal sounds. Validated on a large-scale realistic dataset, it tackles two aspects crucial in the development of automated auscultation analysis: noise and signal-mapping.

The proposed framework addresses the need for improved lung sound quality by using noise-suppression techniques suitable for auscultation applications. It tackles various noise-sources including ambient noise, signal artifacts, patient-intrinsic maskers (heart-sounds, crying); and explores the use of a rich biomimetic feature-mapping that covers the intricate spectro-temporal details of lung sounds, and yields a notable improvement in distinguishing normal/abnormal events when compared to state-of-the-art systems, that tend to fixate on specialized pathologies and global features.

Crucially, this system is further validated on a large patient dataset acquired in the field under realistic clinical conditions. The use of such validation data highlights an additional aspect of the analysis; notably the need for full-patient decisions. Previous studies commonly propose methods for localized interpretations on limited pre-segmented breaths; this entails restricted real-life applicability since it requires a pre-segmentation process that is extremely challenging. Instead, this study hopes to take a step towards realistic applicability of computer-aided diagnosis. In lieu of breath-aligned signal parsing, a short analysis-window is recommended for capturing the manifestation of adventitious sounds of interest while avoiding fixation to highly transient events. A number of challenges remain to be addressed including establishing the association between auscultations and other clinical markers; identifying overlapping non-pathological sounds which can incur significant false positives; and calibrating analysis-windows with respiratory cycles which can benefit the interpretation of the observed patterns.

Acknowledgment

The authors would like to thank the PERCH study group for guidance throughout the completion of this work, and to the patients and families enrolled in this study. The authors would also like to specially thank Dr. J. E. West, who provided invaluable insights about the entire analysis and facilitated the data collection.

References

[1] R. Laennec and W. Osler, *De l'auscultation mediate: Traite du diagnostic des maladies des poumons et du coeur*, vol. 1. Paris, France: Brosson et Chaude, 1819.

[2] L. Liu *et al.*, "Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals," *Lancet*, vol. 388, no. 10063, pp. 3027–3035, 2017.

[3] S. Reichert *et al.*, "Analysis of respiratory sounds: State of the art," *Clin. Med. Circulatory Respiratory Pulmonary Med.*, vol. 2, pp. 45–58, 2008.

[4] B. Flietstra *et al.*, "Automated analysis of crackles in patients with interstitial pulmonary fibrosis," *Pulmonary Med.*, no. 2, p. 1, 2011.

[5] D. Emmanouilidou and M. Elhilali, "Characterization of noise contaminations in lung sound recordings," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 2551–2554.

[6] N. Q. Al-Naggar, "A new method of lung sounds filtering using modulated least mean square adaptive noise cancellation," *J. Biomed. Sci. Eng.*, vol. 6, pp. 869–876, 2013.

[7] K. K. Guntupalli *et al.*, "Validation of automatic wheeze detection in patients with obstructed airways and in healthy subjects," *J. Asthma*, vol. 45, no. 10, pp. 903–907, 2008.

[8] J. Li and Y. Hong, "Wheeze detection algorithm based on spectrogram analysis," in *Proc. 2015 8th Int. Symp. Comput. Intell. Des.*, 2015, vol. 1, pp. 318–322.

[9] D. Emmanouilidou *et al.*, "Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in developing countries," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 9, pp. 2279–2288, Sep. 2015.

[10] S. B. Patel *et al.*, "An adaptive noise reduction stethoscope for auscultation in high noise environments," *J. Acoust. Soc. Amer.*, vol. 103, no. 5, pp. 2483–2491, May 1998.

[11] A. Poreva *et al.*, "Application of bispectrum analysis to lung sounds in patients with the chronic obstructive lung disease," in *Proc. 2014 IEEE 34th Int. Conf., Electron. Nanotechnol.*, 2014, pp. 306–309.

[12] M. Lozano *et al.*, "Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode decomposition and instantaneous frequency," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 2, pp. 486–497, Mar. 2016.

[13] G. Nelson and R. Rajamani, "Accelerometer-based acoustic control: Enabling auscultation on a black hawk helicopter," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 2, pp. 994–1003, Apr. 2017.

[14] R. M. Rady *et al.*, "Respiratory wheeze sound analysis using digital signal processing techniques," in *Proc. 2015 7th Int. Conf. Comput. Intell., Commun. Syst. Netw.*, 2015, pp. 162–165.

[15] N. Nakamura *et al.*, "Detection of patients considering observation frequency of continuous and discontinuous adventitious sounds in lung sounds," in *Proc. 2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 3457–3460.

[16] O. S. Levine *et al.*, "The pneumonia etiology research for child health project: A 21st century childhood pneumonia etiology study," *Clin. Infectious Dis.*, vol. 54, pp. S93–S101, Apr. 2012.

[17] W. H. Organization, *Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Illnesses With Limited Resources*. Geneva, Switzerland: WHO Press, 2005.

[18] E. D. McCollum *et al.*, "The characteristics and reliability of pediatric digital lung sound examinations in six African and Asian countries participating in the Pneumonia Etiology Research for Child Health (PERCH) project," *Amer. J. Respiratory Crit. Care Med.*, vol. 193, 2016, Art. no. A3043.

[19] A. R. A. Sovijarvi *et al.*, "Standardization of computerized respiratory sound analysis," *Eur. Respiratory Rev.*, vol. 10, no. 77, 2000, Art. no. 585.

[20] F. Ghaderi *et al.*, "Localizing heart sounds in respiratory signals using singular spectrum analysis," *Biomed. Eng.*, vol. 58, no. 12, pp. 3360–3367, Dec. 2011.

[21] J. Gnitecki *et al.*, "Qualitative and quantitative evaluation of heart sound reduction from lung sound recordings," *Biomed. Eng.*, vol. 52, no. 10, pp. 1788–1792, Oct. 2005.

[22] D. Flores-Tapia *et al.*, "Heart sound cancellation based on multiscale products and linear prediction," *Biomed. Eng.*, vol. 54, no. 2, pp. 234–243, Feb. 2007.

[23] J. Pesquet *et al.*, "Time-invariant orthonormal wavelet representations," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 1964–1970, Aug. 1996.

[24] P. Basu *et al.*, "A nonparametric test for stationarity based on local Fourier analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 3005–3008.

[25] L. L. LeGasse *et al.*, "Assessment of infant cry: Acoustic cry analysis and parental perception," *Mental Retard. Develop. Disabil. Res. Rev.*, vol. 11, no. 1, pp. 83–93, 2005.

[26] Y. Kheddache and C. Tadj, "Acoustic measures of the cry characteristics of healthy newborns and newborns with pathologies," *J. Biomed. Sci. Eng.*, vol. 06, no. 08, pp. 796–804, 2013.
[27] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *J. Acoust. Soc. Amer.*, vol. 54, pp. 1496–1516, 1973.
[28] J. H. Hospital *et al.*, *The Harriet Lane Handbook: Mobile Medicine Series - Expert Consult*, 19th ed. Philadelphia, PA, USA: Elsevier Mosby, 2011.
[29] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[30] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.
[31] J. Ma *et al.*, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, May 2009.
[32] *American National Standard Methods for Calculation of the Speech Intelligibility Index*, A. S3.5-1997, 1997.
[33] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2224–2237, 2005.
[34] D. Emmanouilidou *et al.*, "A multiresolution analysis for detection of abnormal lung sounds," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 3139–3142.
[35] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, 2005.
[36] X. Lu and M. Bahoura, "An integrated automated system for crackles extraction and classification," *Biomed. Signal Process. Control*, vol. 3, no. 3, pp. 244–254, Jul. 2008.
[37] L. Zhenzhen *et al.*, "A novel method for feature extraction of crackles in lung sound," in *Proc. 5th Int. Conf. Biomed. Eng. Informat.*, 2012, pp. 399–402.
[38] R. Palaniappan and K. Sundaraj, "Respiratory sound classification using cepstral features and support vector machine," in *Proc. IEEE Recent Adv. Intell. Comput. Syst.*, 2013, pp. 132–136.
[39] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[40] F. Jin *et al.*, "New approaches for spectro-temporal feature extraction with applications to respiratory sound classification," *Neurocomputing*, vol. 123, pp. 362–371, 2014.