Rich Representation Spaces: Benefits in Digital Auscultation Signal Analysis

Dimitra Emmanouilidou
Electrical & Computer Engineering Department
The Johns Hopkins University
Baltimore, MD, USA
dimitraem@jhu.edu

Mounya Elhilali
Electrical & Computer Engineering Department
The Johns Hopkins University
Baltimore, MD, USA
mounya@jhu.edu

***Abstract*—When developing automated techniques for analysis of auscultation signals, the choice of a proper representational space that characterizes all attributes of interest in the signal is of paramount importance. In this paper, we investigate different feature representation methods and their benefits in distinguishing auscultation sounds. The importance of choosing an appropriate feature space is explored and validated using trained classifiers that distinguish between normal and abnormal respiratory sounds. Findings of this study are two-fold: i) an increased dimensionality in the feature space can provide a more complete and distinct representation of the delicate breath sounds and ii) dimensionality of the feature space alone is not enough to fully capture discriminative attributes: an informative feature space is even more crucial for extracting accurate, disease-specific characteristics of respiratory sounds.**

Keywords—biomedical signal processing; digital auscultation; lung sounds; respiratory sounds; computerized analysis; multiresolution representation; spectrotemporal analysis; time-frequency analysis; space analysis

I. INTRODUCTION

Chest auscultation has been used over the last 200 years to listen to internal sounds originating from the body and lungs, for diagnostic purposes. In auscultation care, the stethoscope remains the key examination tool, and highly skilled medical personnel are required for the interpretation of the captured body sounds. Over the last few decades, electronic stethoscopes and computer-aided auscultation have improved and facilitated the administration of healthcare across the globe and offered large diversity in clinical training [1][2]. The medical and scientific world has long seen the benefits of computer-aided healthcare, dedicating large numbers of studies to continuous development and improvement of computerized auscultation analysis, with an extended focus on extracting informative features from the raw auscultated sounds.

Lung sound components typically span the range of 50-2500 Hz. Wheezes (100-2500 Hz) and crackles (100-500 Hz) are the most commonly addressed sounds in the literature; adventitious sounds are frequently analyzed in the context of computerized lung sound analysis, using various techniques for discriminating and assessing patients' isolated breath sounds [3]: spectral methods and variations of the Fourier Transform (FT) [4, 5]; and spectro- temporal methods, including the short-time FT (STFT) [6, 7], Mel-scale Frequency Cepstral Coefficients (MFCCs) [8, 9], the Wavelet Transform (WT) and other multi-resolution methods [10, 11]. The plethora of techniques is an indication that many and different features can be used as discriminatory agents in breath sound processing; and among all available methods, the feature representation of choice should be one that adequately captures the characteristics of the breath sounds of interest. But how does the choice of a particular feature dimension affect diagnostic results on breath-sound discrimination? Can this choice be influenced by the particular preference of a feature space over another? These aspects have not been extensively addressed in the literature.

Driven by the need for computationally inexpensive algorithmic methods, in this work we focus on understanding the merit of using an extended feature space, when compared to a richer and more accurate representation. Section II discusses the benefits of compact signal representations when processing and diagnosing auscultated sounds. Feature extraction methods used are described in section III, and findings are presented in section IV.

II. REDUCING DIMENSIONALITY OF EXTRACTED FEATURES

A. *Efficient Computerized Analysis*

Computerized methods for digitally analyzing lung sound signals have not yet been standardized; and, as current standards suffer from low diagnostic accuracy, many recent studies have emerged, demonstrating the need for extended standardized protocols for diagnosing respiratory conditions: in diagnosing pneumonia cases, the gold standard consists of the World Health Organization (WHO) guidelines. In absence of alternative standardized protocols, health care providers all over the world follow these guidelines during their everyday practice. However, a recent study, conducted in a US-based pediatric emergency department, reports that the WHO guidelines merely achieve a 34% accuracy in predicting radiographic pneumonia [12]. The need for alternative protocols is clearly depicted in these accuracy numbers, and computerized analysis is a strong candidate.

A desired standardization of computerized protocols does not only hope to improve health care administration, but will

Research supported by Grants NIH 1R01AG036424, ONR N000141010278, N000141612045 and N000141210740.

2374-7390/16 $31.00 © 2016 IEEE
DOI 10.1109/SiPS.2016.20

also help resolve variations of terminology among healthcare professionals and medical publications and researchers [13]. And as auscultation sound analysis awaits standardization, sophisticated signal processing techniques are being built with advanced diagnostic capabilities, where the use of appropriate, low dimensional features will help extract crucial information and more accurately, and will be able to aid physicians administer better health care with faster diagnostic procedures, around the globe. And if a faster health care administration doesn't seem a matter of real concern, according to a recent study [14]: an average physician would need to spend about 22 hours per day to provide the recommended care to every one of her patients. The objective of extracting suitable and compact signal characteristics will significantly help improve processing results as well as decision-making time.

III. Methods

In computerized lung sound analysis, various techniques for extracting the relevant and crucial information from the sound signal have been studied in the literature. Extracted signal characteristics vary from low to high-dimensional spaces and from moderate to rich representations. Here, we explore 7 methods for extracting lung sound characteristics with a varying size of the feature space. We examine each one of these representations and explore their ability to efficiently distinguished normal from abnormal breath sounds, by imploring SVM classifiers.

A. Data and Annotations

Trained medical personnel digitally recorded lung sound from children, ages 1 to 59 months (average age 11 ± 11.43), in outpatient or busy clinical settings, in Zambia, Kenya, Gambia, South Africa, Bangladesh, and Thailand PERCH sites [15]. The auscultation protocol called for 7 s recordings over 8 body sites with a digital ThinkLabs Inc. stethoscope, sampling at 44.1 kHz. In total, 1157 sick children were enrolled as *cases,* having WHO-defined severe or very severe pneumonia, or *controls*, without clinical pneumonia.

A standardized panel of 8 trained physicians interpreted the recordings and indicated respiratory findings in 8 body locations. A refined label was given for each location (site), corresponding to a clip of arbitrary length that best represented findings. Labeled clips were included in the study only if there was agreement among primary listeners on a conclusion of wheeze (a site with wheezing breaths) or normal (a site with no wheeze or crackle sounds). The refined annotations were split into non-overlapping 2-sec segments, and grouped into: *Normal*, containing normal-annotated breath sounds (breaths without a wheeze annotation) and *Abnormal*, containing wheeze annotations. In total, 935 *Abnormal* and 1231 *Normal* intervals were isolated.

B. Pre-processing

A 4^{th} order low-pass Butterworth filter at 4 kHz cut-off was applied to all recorded breath sounds before resampling at 8 kHz and normalizing to zero mean and unit variance. As lung sound content is found below 4 kHz [3], no loss of information was anticipated after resampling.

All lung sound recordings were acquired in busy clinics and were highly challenged by various noise contaminations. Contaminations included environmental noise, such as vehicle sounds, children crying in the waiting room or phones ringing, and subject-centric noises coming from infants being restless or crying during examination. In order to achieve a cleaner lung sound signal before continuing with further processing, we invoked a spectral subtraction algorithm efficiently tuned to breath sounds, validated for suppressing ambient noise while preserving the delicate breath content [16]. Although environmental noise was highly suppressed by the algorithms, the majority of the recordings were still found to be contaminated by non-breath sounds, including subject's cry reverberation sounds and electronic or stethoscope noise which frequently occurred due to the challenging young age of the enrolled patients and the busy clinics. The remaining ambient contaminations can have notably overlapping profiles with the lung sound content along both time and frequency [17], impeding further analysis.

C. Feature Extraction Methods

All data coming from the refined annotations were used for feature extraction. One of the most common processing techniques used is the frequency (spectrum) analysis. We invoked the method described by Waitman et al. for acquiring a binned version of the signal's Fourier representation within the frequency range of interest [4], which we call here FFTW; the predefined spectrum bin size varied in $\{100, 200\}$.

Next we extracted MFCCs, capturing information both along the time axis and a transformed frequency. They encode information about the peak energies or resonances of a sound signal and we can consider them here as indirectly related to the impulse response of a system related to the thoracic area. Different formations of the thoracic area are expected to yield changes in the MFCC sequences of the recorded sound. Here we invoke the method described in [8], called here MFCCJ. Ten or twenty triangular filters were used to extract the MFCCs, over 35 msec frames with 50% overlap, resulting in 100 or 200 feature dimensions.

The next set of features comes from a multi-dimensional representation, inspired by the way sound is being processed in the auditory pathway. Briefly, a bank of 128 cochlear filters $h(t;f)$, modeled as constant-Q asymmetric bandpass filters equally spaced on a logarithmic frequency scale spanning 5.3 octaves, transforms the sound signals $s(t)$ into a modified short-time spectral representation representing the sharpened response of auditory nerve signals. A midbrain model is achieved using short term integration (or low-pass operator $\mu(t;\tau)$ with constant τ=2msec), resulting in a final time frequency representation, the auditory spectrogram (1).

$$y(t,f)=max[\partial_f(\partial_t(s(t) * h(t;f))),0]*\mu(t;\tau). \quad (1)$$

The next step models the processing signals undergo at the central auditory stages where a rich representation is obtained. The operation is modeled as 2D affine Wavelet transform. Each filter is tuned (Q=1) to a specific temporal modulation ω_0 (or rate in Hz) and spectral modulation Ω_0 (or scale in cycles/octave or c/o), as well as directional orientation in time-frequency space (+ for upward and − for downward). The response of each cortical neuron is given by

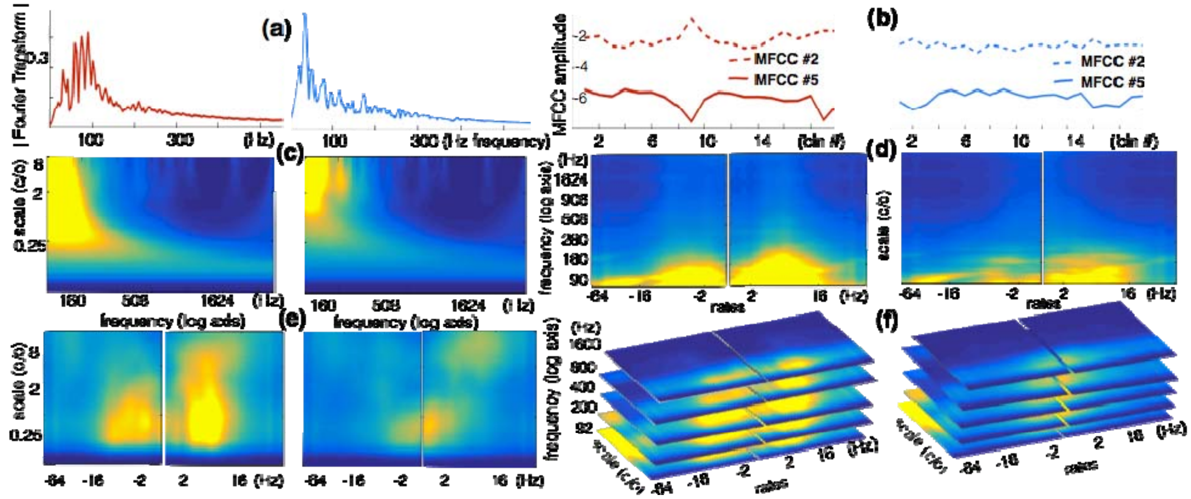


Fig. 1. Different feature extraction representations of an Abnormal (left column) and a Normal (right column) breaths recorded from two distinct enrolled subjects: a) FFTW; b) MFCCJ; c) SF; d) RF; e) SR; f) SRF.

$$r_{\pm}(t,f;\omega_0,\Omega_0)= y(t,f) *_{t,f} STRF_{\pm}(t,f;\omega_0, \Omega_0) \qquad (2)$$

where $*_{t,f}$ corresponds to convolution in time and frequency and $STRF_{\pm}$ is the 2D filter response of each cortical neuron. The resulting cortical representation is a mapping of the time waveform onto a high-dimensional space. 28 scale filters in 0-8 c/o and 21 directional rate filters were used in 8-128 Hz, in logarithmic steps [10, 18]. We isolated 5 multi-resolution representations derived from the cortical features:

1) Feature set SF, corresponding to the spectral modulations found in the breath sounds, along the frequency axis: $SF(f;\Omega_0)=\int_t y(t,f) *_{t,f} STRF_{\pm}(t,f;\Omega_0)dt$;

2) Feature set RF, corresponding to the temporal variations found in the breath sounds, along the frequency axis: $RF(f;\omega_0)=\int_t y(t,f) *_{t,f} STRF_{\pm}(t,f;\omega_0)dt$;

3) Feature set [SF, RF] corresponding to the concatenation of feature sets SF and RF above;

4) Feature set SR, corresponding to the spectral and concurrent temporal modulations found in the breath sounds, disregarding the information of the frequency axis: $SR(\omega_0, \Omega_0)=\int_f \int_t y(t,f) *_{t,f} STRF_{\pm}(t,f; \omega_0, \Omega_0)dt\, df$;

5) Feature set SRF described in (2), integrated over time: $SRF(f; \omega_0,\Omega_0) = \int_t y(t,f) *_{t,f} STRF_{\pm}(t,f; \omega_0,\Omega_0)dt$. This representation is rich in information, containing concurrent temporal and spectral modulations, along the frequency axis

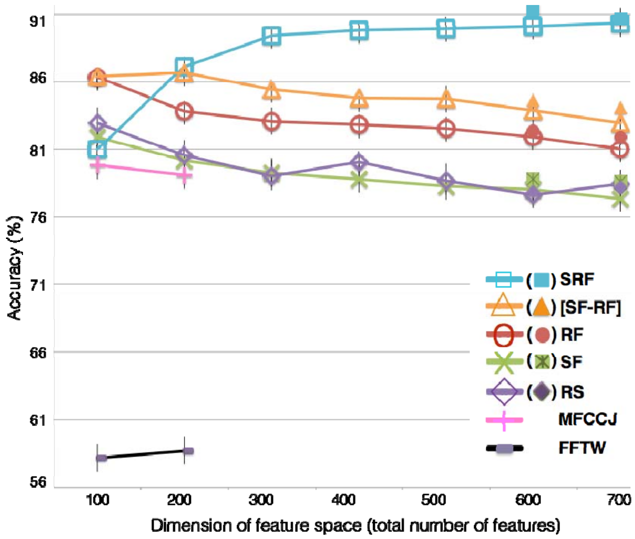
All sets contain rich information extracted from the breath sounds: information on how fast or slow the particular frequency contents change and in which directionality (RF) or information on how wide- or narrowband the breath content is, along frequency (SF). SRF provides a high dimensional representation (original dimensions 28 x (2x21) x 128) of concurrent spectral and temporal modulations along the frequency axis. Data dimensionality was achieved using tensor-Singular Value Decomposition (SVD). Data were unfolded along each feature dimension and the principal components were calculated from the covariance matrix. Components were ranked and selected according to their ability to capture the total feature variance.

IV. Results

The time waveforms of the lung sounds were augmented and transformed into richer representations using the various methods discussed. Examples of how each feature space transforms the breath sounds are shown in Fig.1.

The search for an outperforming classifier is not in of the scope of this paper. In this work, Binary Support Vector Machines (SVM) classifiers were used for the two-class data discrimination problem, with Radial Basis Function kernels (γ=0.1). Classifiers were trained on 90% of the data and tested on 10% of the data, using a 10-fold cross validation and 10 independent Monte Carlo (MC) runs, for each different feature set and feature set dimension. To avoid biasing the classifiers, an equal size of Abnormal and Normal sounds were used for every cross-validation, by randomly sampling the Normal group to match the size of the Abnormal group. Performance was reported in terms of Accuracy, where Accuracy=100*(True Positives + True Negatives)/(all). The discriminative capabilities of the different feature sets are illustrated in Fig.2, where the x-axis depicts the effect of the feature set size on discriminatory accuracy. Accuracy values appear with their standard deviation errors across MC runs. For the multi-resolution representations, the number of features corresponds to the number of principal components, ranked and selected to capture at least 99% of the variance.

Evidently, space representations that are less rich do not seem to capture distinguished features in an efficient way:



Fig, 2. Accuracy curves on the task of discriminating Normal from Abnormal breath sounds. The size of feature space varies along x-axis.

method FFTW performed in the low side of the accuracy map; as the data representation was enriched using the MFCCJ method, performance results increased. The increment in the dimensionality for FFTW and MFCCJ methods did not seem to affect discrimination.

Looking at various multiresolution representations, methods SF, RS and RF all proved to adequately capture lung sound characteristics and achieve discrimination accuracy results above 76%. The error bar levels indicate that although these set of features capture different characteristics in nature, they can perform equally well to the task at hand; a small number of singular vectors were enough to capture most of the data variation. A highly reduced space of 100 total features was adequate here for capturing distinct sound characteristics. SF, RS and RF sets enclose unique information and were all shown capable on the discrimination task. Similarly we expect the concatenated [SR-RF] set to achieve equal or better performance. Looking at Fig. 2, as expected, [SR-RF] moderately surpasses the individual marginal spaces, revealing that both spectral and temporal modulations of the signal capture unique, necessary lung sound content. Further, notice how in the [SR-RF] space, increased dimensionality appears to introduce a slight confusion to the classifier. This can be attributed to a number of reasons: First, the corresponding error bars show a fair performance variation and might partially explain the apparent performance inconsistency with increased dimension. The inherent irregular and unpredictable ambient noise of the breath sounds can also be a contributing factor, occurring unexpectedly, and confusing the classifier both with relation to time and spectral information, when increased feature details are included [17]. Similarly the inclusion of more dimensions can also signify an increase in the amount of included shared information between groups. A final prospective factor is the well-known curse of dimensionality; increased signal representation spaces require an increased number of training examples, with practically an exponential relation. Therefore it is not surprising to experience fluctuations in the performance curve, as the feature dimensions increase and the number of training examples remains constant [19]. Part of this confusion seems to disappear when an even richer and improved representation space is used: the SRF space captures simultaneous modulations along the spectral and temporal axis, and provides a highly informative space and a more robust representation. With a 99.87% reduction of the original space (when total feature size is reduced to 200) the classifier achieves favorable accuracy results above 86%.

To further explore the apparent decrease of the accuracy slope for higher dimensions, we created a second pool of data, reducing the *Normal* group to 946 sounds, where we excluded breaths highly corrupted by loud crying or prominent electronic noise or recording interruptions. Due to the significant overlap of the abnormal breath profiles with noise contamination [16], we did not exclude sounds from the *Abnormal* group but randomly matched the number of Normals during classification, to avoid bias. The same setup was used, including the various feature extraction methods and SVM classification. The obtained results were similar to the ones presented in Fig.2, besides a modest overall increase in accuracy levels. We include results for dimension={600,700}, showed with filled color markers in Fig.2 (see corresponding legend marker in parenthesis). Despite the use of this “cleaner” pool of data, there remained an apparent drop in the accuracy slope for most of the multi-resolution feature sets. The exclusion of prominently corrupted breath sounds did not change the performance curves, indicating that a) the nature of these particular contaminations did not previously affect the classifier’s capabilities and the inherent discrimination difficulty comes from complex or convoluted environmental noise, or from the overlapping shared information among breath sounds of the two groups, or b) exploring detailed temporal and spectral information of the breath manifestations have introduced extra confusion and the benefits of a high-dimensional space are obscured by its complexity.

V. Conclusion

The choice of the most appropriate feature set for data classification is considered one of the holy grails in machine learning. This feature set has to take into account the commonalities and differences within and across classes, and be robust to a number of factors (noise, unpredictable variability). In the case of sound analysis and classification, a combination of spectral and temporal features has often been sought to represent signals along their frequency content and dynamics as sound evolves over time.

Overall, the results demonstrate that a rich space is required to best capture the intricate details inherent in lung sounds, particularly for the purpose of distinguishing normal breathing patterns from abnormal ones. A rich space is necessary but not sufficient. Even with fixed dimensionality of the feature space, a representation that combines *joint* spectral and temporal attributes of the signal (SRF) is more informative than one that combines the spectral *and* temporal features separately [SF-

RF]. These joint modulations that change along *both* time and frequency (e.g. a frequency modulation) cannot be easily captured by the marginal spectral and temporal distributions; even if both representations span a rich feature space of few hundred dimensions. That being said, the richness of the feature space also faces its own curse of dimensionality. As the number of dimensions was increased, it does not always lead to an increase in classification accuracy. Further constraints and appropriate prior knowledge on the signal characteristics can be more important than dimensionality.

Due to the challenging nature of the dataset (pediatric sounds acquired in busy and outpatient clinics), a certain degree of noise or distortion is unavoidable. To what extent the inherent noise affects the observed accuracy results, as opposed to more pathological reasons, remains to be seen. These findings merit further exploration in order to expand the possibilities of automated lung sound analysis that can be deployed in the field without limitations and constraints on environmental or pathological requirements. The current study aims to emphasize the role of a rich and carefully carved feature space as a necessary step in developing such unconstrained auscultation systems. Continuing work on both well-controlled and challenging recordings will help determine and isolate the direct effect of the inherent noise, as opposed to the effect of the shared content between normal and abnormal breaths. Extended work with an augmented pool of adventitious sounds will help medical scientists understand the value of different feature representations with respect to the auscultated sounds of interest.

ACKNOWLEDGMENT

The authors would like to thank Dr. Eric D. McCollum, Johns Hopkins School of Medicine and Daniel E. Park, Johns Hopkins Bloomberg School of Public Health, for providing the lung sounds, part of the PERCH study [15].

REFERENCES

[1] C. Hou, Y. Chen, L. Hu, C. Chuang, Yu-Hsien Chiu and Ming-Shih Tsai, "Computer-aided auscultation learning system for nursing technique instruction," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE,* 2008, pp. 1575-78.

[2] A. Gurung, C. G. Scrafford, J. M. Tielsch, O. S. Levine and W. Checkley, "Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis." *Respir. Med.,* vol. 105, pp. 1396-1403, Sep, 2011.

[3] A. R. A. Sovij\arvi, J. Vanderschoot and J. E. Earis, "Standardization of computerized respiratory sound analysis," *European Respiratory Review,* vol. 10, pp. 585, 2000.

[4] L. R. Waitman, K. P. Clarkson, J. A. Barwise and P. H. King, "Representation and classification of breath sounds recorded in an intensive care setting using neural networks." *J. Clin. Monit. Comput.,* vol. 16, pp. 95-105, 2000.

[5] L. J. Hadjileontiadis, *Lung Sounds: An Advanced Signal Processing Perspective.* San Rafael, California: Morgan & Claypool Publishers, 2009.

[6] B. A. Reyes, S. Charleston-Villalobos, R. González-Camarena and T. Aljama-Corrales, "Assessment of time–frequency representation techniques for thoracic sounds analysis," *Comput. Methods Programs Biomed.,* vol. 114, pp. 276-290, 5, 2014.

[7] R. J. Riella, P. Nohama and J. M. Maia, "Method for automatic detection of wheezing in lung sounds." *Brazilian Journal of Medical and Biological Research,* vol. 42, pp. 674-684, jul, 2009.

[8] Jen-Chien Chien, Huey-Dong Wu, Fok-Ching Chong and Chung-I Li, "Wheeze detection using cepstral analysis in gaussian mixture models," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE,* 2007, pp. 3168-3171.

[9] P. Mayorga, D. Ibarra, V. Zeljkovic and C. Druzgalski, "Quartiles and mel frequency cepstral coefficients vectors in hidden markov-gaussian mixture models classification of merged heart sounds and lung sounds signals," in *High Performance Computing & Simulation (HPCS), 2015 International Conference on,* 2015, pp. 298-304.

[10] D. Emmanouilidou, K. Patil, J. West and M. Elhilali, "A multiresolution analysis for detection of abnormal lung sounds," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 2012, pp. 3139-3142.

[11] M. Najafian, D. Irvin, Y. Luo, B. S. Rous and J. H. L. Hansen, "Employing speech and location information for automatic assessment of child language environments," in *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE),* 2016, pp. 1-5.

[12] S. Wingerter, R. Bachur, M. Monuteaux and M. Neuman, "Application of the World Health Organization criteria to predict radiographic pneumonia in a US-based pediatric emergency department," *Pediatric Infect Dis J,* vol. 31, pp. 561-564, 2012.

[13] R. L. Wilkins, J. R. Dexter, R. L. H. Murphy and E. A. DelBono, "Lung Sound Nomenclature Survey," *Chest,* vol. 98, pp. 886-889, 10, 1990.

[14] J. Altschuler, D. Margolius, T. Bodenheimer and K. Grumbach, "Estimating a reasonable patient panel size for primary care physicians with team-based task delegation," *Annals of Family Medicine,* vol. 10, pp. 396—400, 2012.

[15] O. S. Levine, K. L. O'Brien, M. Deloria-Knoll, D. R. Murdoch, D. R. Feikin, A. N. DeLuca, A. J. Driscoll, H. C. Baggett, W. A. Brooks, S. R. Howie, K. L. Kotloff, S. A. Madhi, S. A. Maloney, S. Sow, D. M. Thea and J. A. Scott, "The Pneumonia Etiology Research for Child Health Project: a 21st century childhood pneumonia etiology study," *Clin. Infect. Dis.,* vol. 54 Suppl 2, pp. S93-101, Apr, 2012.

[16] D. Emmanouilidou, E. D. McCollum, D. E. Park and M. Elhilali, "Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in developing countries," *IEEE Trans Biomed Eng,* vol. 62, pp. 2279-2288, 2015.

[17] D. Emmanouilidou and M. Elhilali, "Characterization of noise contaminations in lung sound recordings," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* Osaka, Japan, 2013, pp. 2551-2554.

[18] T. Chi, P. Ru and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J Acoust Soc Am,* vol. 118, pp. 887-906, 2005.

[19] D. Donoho, "High-dimensional data analysis : The curses and blessings of dimensionality," 2000.