Biomedical Signal Processing and Control 85 (2023) 104852

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Reference free auscultation quality metric and its trends

Annapurna Kala [a], Eric D. McCollum [b], Mounya Elhilali [a,*]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA
[b] Global Program of Pediatric Respiratory Sciences, Eudowood Division of Pediatric Respiratory Sciences, Department of Pediatrics, Johns Hopkins School of Medicine, Baltimore, USA

ARTICLE INFO

Keywords:
Auscultation quality
Computerized auscultation analysis
Clinical validation
No-reference metric

ABSTRACT

Stethoscopes are used ubiquitously in clinical settings to ‘listen’ to lung sounds. The use of these systems in a variety of healthcare environments (hospitals, urgent care rooms, private offices, community sites, mobile clinics, etc.) presents a range of challenges in terms of ambient noise and distortions that mask lung signals from being heard clearly or processed accurately using auscultation devices. With advances in technology, computerized techniques have been developed to automate analysis or access a digital rendering of lung sounds. However, most approaches are developed and tested in controlled environments and do not reflect real-world conditions where auscultation signals are typically acquired. Without a priori access to a recording of the ambient noise (for signal-to-noise estimation) or a reference signal that reflects the true undistorted lung sound, it is difficult to evaluate the quality of the lung signal and its potential clinical interpretability. The current study proposes an objective reference-free Auscultation Quality Metric (AQM) which incorporates low-level signal attributes with high-level representational embeddings mapped to a nonlinear quality space to provide an independent evaluation of the auscultation quality. This metric is carefully designed to solely judge the signal based on its integrity relative to external distortions and masking effects and not confuse an adventitious breathing pattern as low-quality auscultation. The current study explores the robustness of the proposed AQM method across multiple clinical categorizations and different distortion types. It also evaluates the temporal sensitivity of this approach and its translational impact for deployment in digital auscultation devices.

1. Introduction

Lung sounds have been used for the diagnosis of pulmonary diseases for centuries since the invention of stethoscopes [1]. With advances in digital sensing and analysis technologies, Computerized Auscultation Analysis (CAA) is becoming more popular and opening new frontiers for telemedicine, automated diagnostics, and versatile healthcare [2, 3]. Rapid progress in deep learning techniques have revolutionized a number of fields from computer vision to audio system [4–6], and have contributed to important breakthroughs in computerized analysis of adventitious lung sounds as pathological indicators [7–11]. Access to digital auscultations is also facilitating the incorporation of lung sound recordings into electronic health records and long-term monitoring and longitudinal analysis of health data [12]. Such data is paving the way for new possibilities for continued monitoring and further mining of biomarkers and pulmonary sounds using deep learning technologies.

One of the limiting factors hampering progress in the field of mining digital auscultations is data access and curation. The issue of data is further compounded by the complexity and variability of clinical settings that stem from the clinical environment, the devices used, as well as the training of the user. When a physician or healthcare worker uses a stethoscope to ‘listen’ to breathing patterns from the lung, much of the ambient environment is being picked up by the sensor of the device. While a physician relies on their medical training to ignore these ambient sounds and focus on the breathing patterns, a computerized system requires additional processing in order to properly access these unadulterated lung sounds. Despite technological advancement, data curation remains a major bottleneck particularly in the field of digital auscultation. In addition to challenges commonly faced with developing machine learning tools for medical screening and diagnostics, the field of sound auscultation poses unique hurdles with regard to the very nature of the signals acquired. One of the main goals of ‘listening’ to body sounds is to identify abnormal breathing patterns that are indicative of pathological conditions such as pneumonia or bronchiolitis. These abnormal lung sounds such as wheezes (long whistling sounds) and crackles (series of short explosive sounds) often share spectrotemporal characteristics that are very similar to ambient

[*] Corresponding author.
E-mail address: mounya@jhu.edu (M. Elhilali).

https://doi.org/10.1016/j.bspc.2023.104852
Received 3 December 2022; Received in revised form 21 February 2023; Accepted 11 March 2023
Available online 23 March 2023

noise making the two easily confusable or causing masking of abnormal patterns by background noise.

Under controlled-environments, a number of studies have shown that adventitious lung sounds have well understood properties that can be used for screening or diagnosis of specific lung pathologies [13–15]. However, auscultations collected in busy clinical settings tend to show a great deal of variability depending on the ambient conditions at the time of recording [16]. Moreover, lung sound recordings are collected in a wide range of clinical settings which can induce variability due to differences in setting (physician's office, ER, rural clinic, etc.), differences in devices and sensors, and the temperament of patients especially when dealing with infants. This variability ultimately results in a great deal of inconsistencies in auscultation quality. When an expert ear is present (healthcare worker, physician), they can evaluate this quality on the spot and make adjustments when possible (move the position of stethoscope on the body, calm the patient down, close a door if needed). However, this human interpretation can be a limiting factor in order to curate data for automated processing and development of diagnostic technologies establishing a need for auscultation standardization [17]. The issue of the standardization of data quality can be a stumbling block in order to automate both the acquisition as well as preprocessing of large amounts of data, both for development of learning algorithms as well as potential use for medical records and long-term tracking.

One of the challenges facing the issue of data standardization is that there is no agreed-upon definition as to what constitutes "high-quality" data in the domain of digital auscultations. While several denoising algorithms exist to tackle the noise problem, it is important to have a metric to gauge the quality of processed signals. A suitable quality metric is necessary for evaluating denoising algorithms, as well as appraising the utility of the signal in making a final diagnosis. There are a few quality assessments focusing on *heart auscultation* which further emphasize the importance of assessing the quality of the signal used for the disease detection [18–20]. In the case of pulmonary signals, obtaining an effective quality metric comes with additional considerations. Since abnormal lung sounds have similar spectrotemporal properties to ambient noise, the metric should not flag an abnormal lung sound as low quality (or noisy) thereby disposing of critical diagnostic information. In previous work [21], we developed a pulmonary auscultation quality metric derived from select spectral features and data driven features using a *linear* regression model. In the present work, we extend this framework by considering a larger set of highly informative spectro-temporal features informed by feature selection techniques. Moreover, realizing the non-linear trends of the "quality" space, we focus the mapping of a no-reference auscultation metric on a nonlinear transformation of signal properties. We further show how this quality metric follows the expected trends across different view points of clinical assessment like level of agreement, surety, and level of interpretability of the data. In addition, we analyze the robustness of this metric across different types of noises, as well as explore the temporal sensitivity of the metric in flagging a sudden-onset noise, which could assist the user to recollect the auscultation data immediately.

The paper describes the data and proposed framework for the Auscultation Quality Metric (AQM) in Sections 2 and 3. Section 4 analyzes the importance of features included in the algorithm. Section 5 discusses the results of experiments with different environments and ambient settings. The trends in AQM across several clinical viewpoints are presented in Section 6. Section 7 explores the temporal sensitivity of the proposed method relative to transient maskers and Section 8 presents further discussion and implications of the proposed scheme.

2. Data and methods

2.1. Data acquisition and preprocessing

The analysis is based on signals collected by the Pneumonia Etiology Research for Child Health (PERCH) study group [22]. Data is collected using a Thinklabs ds32a digital stethoscope at a rate of 44.1 KHz and acquired at 9 sites spanning over 7 countries (The Gambia, Mali, Kenya, Zambia, South Africa, Bangladesh, and Thailand) during the period between 2011–2014. PERCH lung sounds are often masked by ambient noises expected in pediatric settings such as musical toys, background chatter in the waiting room, children crying, vehicle sirens, and mobile or other electronic interference. Subjects are pediatric patients between 1–59 months old, hence intense crying often contaminates the recordings. All signals are pre-processed by applying a low pass fourth-order Butterworth filter with a cutoff at 4 kHz, downsampled to 8 kHz, centered to zero mean and unit variance, and denoised using a noise-cancellation algorithm to deal with ambient noise, cries, and heart sound contamination [23].

After pre-processing, a panel of eight listening experts (six pediatricians and two pediatric-experienced physicians) assessed the auscultations. In addition to identifying pathological indicators -if any-, the panel provides additional descriptors about each recording: clinical interpretability, presence of crying, presence of signal clipping. A total of 13.3 hours of auscultation sounds (out of 40.42 hours of recordings) are flagged as not clinically interpretable by at least one reviewer. In case two reviewers assigned to a recording mark it as clinically not-interpretable, it is delegated for further arbitration by up to two additional reviewers. The listening panel further annotates interpretable signals as either normal sounds or adventitious lung sounds (containing either wheezing or crackles or both) following their definitions in the American Thoracic Society guidelines, as well as flag segments of intense crying in the recorded auscultations. Additionally, the panelists assign a level of certainty to all annotated normal/abnormal labels by marking them as "definite", "probable" and "non-interpretable" (see [24] for details on annotation methodology).

In addition to these markers, we conduct a second listening panel asking two expert physicians to rate the quality of a recording. The experts evaluate 92 randomly selected lung sound recordings comprising definite interpretable signals as well as noisy versions that are deliberately corrupted with controlled levels of noise using the BBC sound effects database (chatter and crowd noises) [25] at signal to noise ratios of −5, 10, and 20 dB. The rating of the lung sound quality is on a scale of 1 (clinically completely uninterpretable) to 5 (the highest quality). The data comprises 10–20 seconds audio clips.

2.2. Data selection for quality assessment

Segments for which a majority of expert listeners agree on the clinical diagnosis with *high confidence* as normal or abnormal are defined as 'High Quality' database (a total of 11.5 h) of auscultation signals. One of the key aspects of a deployable auscultation quality metric is to not discard the abnormal lung sounds as poor quality while acknowledging they share spectral properties with the ambient noise. To attain this, a subset $\Gamma^{HQ}_{E-Train}$ (40%) of this database with an equal number of normal (no acute lower respiratory infections) and abnormal (acute lower respiratory infection indicators such as wheezes and crackles) cases is used to learn the data driven quality features which capture the 'high-quality' profile. A random subset of the high quality data (20%) $\Gamma^{HQ}_{R-Train}$ is further systematically corrupted with noise from BBC sound effects database [25] and white noise. All the noises are added to the auscultation audio with signal to noise ratios from −5 dB to 40 dB in gradual increments of signal to noise ratio for training the quality metric regression network in a cross-validation fashion. We ensure that equal proportions of BBC ambient sounds, BBC speech sounds, and white noise are used to capture the entire surround noise profile in the Auscultation Quality Metric training. This metric is then tested on the rest of the high quality data (5.52 h). The evaluation data set also includes data with inconclusive agreement among reviewers (7.6 h) as well as 8 hours of auscultation data in which the pediatric patient was noted to be crying.

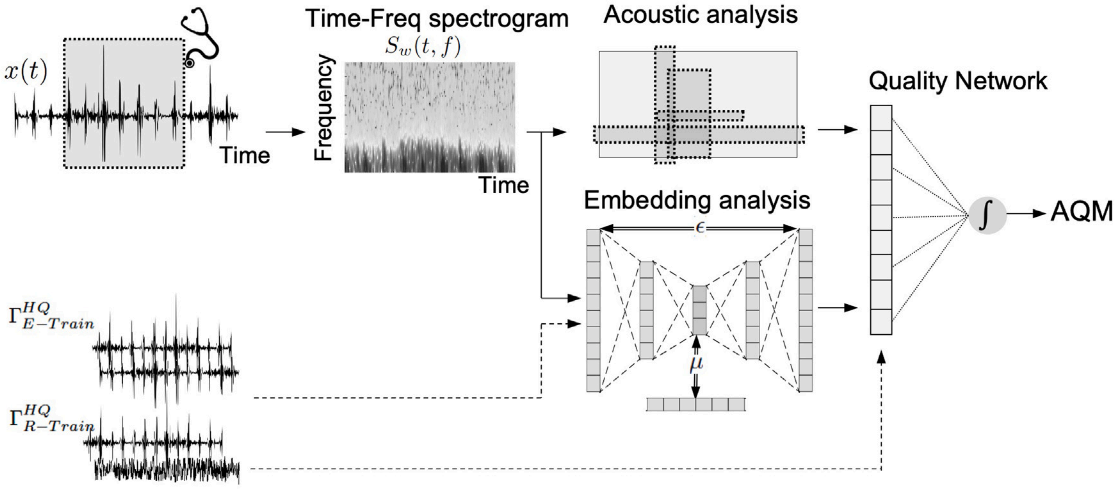


Fig. 1. Overview of the Auscultation Quality Metric algorithm: A short-time analysis of the auscultation waveform is performed over a moving window. Each segment is then converted to an auditory spectrogram followed by two analysis streams: an Acoustic analysis of spectral and temporal signal attributes, and an Embedding analysis of learned higher-level interpretations of the signal space. These attributes are integrated using a non-linear quality network to formulate the final Auscultation Quality Metric.

2.3. Statistical analysis

The quality metric analyses are done on many clinical and carefully perturbed subsets of the entire auscultation data to evaluate its robustness. To this end, we report the mean values and standard errors of quality metric samples of each of these subsets. We then compare each of the samples statistically against a normal clean subset. As the samples failed normality tests, we perform Wilcoxon Rank Sum Test to report the statistical significance of t-tests.

3. Auscultation quality metric (AQM)

Given an auscultation signal $x(t)$, the proposed quality metric performs a quality analysis along two parallel paths (Fig. 1): (1) an acoustic analysis which extracts low-level spectrotemporal profiles of the signal, and (2) an embedding analysis which extracts high-level descriptors of auscultations including normal and pathological patterns. These paths are integrated using non-linear regression.

3.1. Acoustic analysis

The audio waveform $x(t)$ is segmented into W segments $x_1, x_2, \dots, x_W$ using a 2 seconds rectangular moving window with 50% overlap. Each segment is mapped onto a time–frequency spectrographic representation $S_w(t,f)$ as proposed in Chi et al. [26], then are further processed to extract spectral and temporal characteristics of the signal as outlined in Algorithm 1 (see [27] for more details). The specific features of interest are:

- **Average spectral energy ($E[S]$):** is obtained by considering the average adjacent frequency bin energy content in an auditory spectrogram.
- **Scale Average Energy ($E[\hat{S}]$):** represents the average energy spread in the spectrogram over a bank of 28 log-spaced spectral filters parameterized by the spectral modulation Ω ranging between 0.25 and 8 cycles/octave.
- **High Modulation Rate Energy (HR):** reflects the roughness of the signal and is analyzed by averaging the energy content in temporal modulation frequencies ω above 30 Hz.
- **Low Modulation Rate Energy (LR):** is derived from the energy content in temporal modulation frequencies between 1 and 30 Hz.
- **Pitch ($\hat{F}_o$):** is calculated by selecting the best match of the spectral profile of each time slice ($S_w(t_o, f)$) from a set of pitch templates (T_k) and generating a maximum likelihood estimate to fit a pitch frequency (P_k) to the selection [28].
- **Bandwidth (BW):** quantifies the range of frequencies with non-zero content in the signal and is computed as the weighted distance of the spectral profile from its centroid.
- **Spectral Flatness (SF):** reflects the degree of uniformity in the frequency response of a signal and is formulated as the geometric mean of the spectrum divided by its arithmetic mean [29].
- **Spectral Irregularity (SI):** represents the variability in the frequency content of the signal and is calculated as the difference in strength between adjacent frequency channels.

3.2. Embedding features

A four-layer convolutional neural network autoencoder is trained unsupervised on $\Gamma_{HQTrain}$ dataset which is considered clinically highly interpretable to obtain a profile of high quality lung sounds. The network comprises 80 3×3 kernel filters in the first two layers mapping to feature space (encoder) and 80 2×2 kernel filters reconstructing the spectrogram from feature space (decoder). ReLU activations in the network ensure the non-linearity of feature space. This network is trained using Adam optimizer at a learning rate of 0.001. The training dataset $\Gamma^{HQ}_{E-Train}$ having an equal number of normal and abnormal lung sounds ensures that adventitious breathing patterns are acknowledged as 'high-quality' by the network instead of misrepresenting them as poor quality. Once trained, two parameters are extracted from this network as depicted in Fig. 1, and used to supplement the acoustic features:

- **Mean Feature Error (μ):** A dense low dimensional embedding (32 × 32) is obtained by passing the input spectrogram $S_w(t,f)$ (32 × 128 dimensions) through the first two layers of the CNN Autoencoder. An average of all the training embeddings acts as the high-quality 'template' of the auscultation data in the feature space. The L2 distance of the unsupervised features of the test signal from this average feature template is taken as the corresponding Mean Feature Error.
- **Reconstruction Error (ϵ):** Given the Autoencoder is trained on clean data, a good quality lung sound would be closer to high-quality training data and gives better reconstruction. The L2 distance of the reconstruction with the original spectrogram indicates the reconstruction error of the test recording on the high-quality network and acts as the second embedding feature.

Table 1
Mean and Standard Error of ratio of Absolute Error Values before and after feature shuffling to evaluate the importance of each feature.

Shuffled feature	Ratio of AE	P-Value
$E[S]$	16.74 ± 0.095	0
$E[\hat{S}]$	5.90 ± 0.025	$2.6e^{-4}$
HR	10.69 ± 0.080	0
LR	9.76 ± 0.055	$3.4e^{-31}$
$\hat{F}_o$	4.17 ± 0.029	0.013
BW	16.04 ± 0.085	0
SF	3.09 ± 0.020	$1.15e^{-6}$
SI	5.7 ± 0.050	$1.01e^{-8}$
μ	6.6 ± 0.052	$1.75e^{-80}$
ϵ	1.96 ± 0.009	$e{-}3$

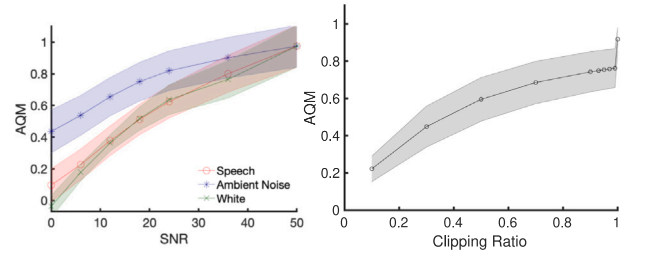
3.3. Quality network

Both signal-centric and learned features (using the autoencoder) are weighted non-linearly to get an overall quality metric. The ten features are first mapped onto a log-scale and are further integrated using multivariate non-linear regression performed by an artificial neural network (ANN) with ten input nodes and one output node, with a sigmoid activation to scale outputs between 0 (clinically uninterpretable) and 1 (perfect quality). To train such a network in a supervised fashion, regression labels for $\Gamma^{HQ}_{R-Train}$ are used ranging from 0 to 1 with 0 assigned to signals corrupted at −5 dB signal-to-noise ratio and 1 to the un-corrupted high-quality lung sounds. The intermediate labels between 0 and 1 gradually reflect increasing levels of signal to noise ratios between −5 dB and 30 dB.

Throughout the analysis presented in this work, we map each 2 seconds signal segment onto an AQM score. For longer signals, AQM scores are calculated using a moving window with 50% overlap, then averaged across all windows of the signal to yield a single score. An alternative approach of averaging the acoustic and embedding features across the entire duration of the signal, and then mapping to a single AQM score yields quantitatively similar results and is not presented here.

4. Feature importance analysis

The choice of AQM features is carefully designed to balance signal profile and informative representations that do not discard adventitious auscultation events. To evaluate the contribution of each feature to the final 'quality metric', we analyze feature importance using a permutation analysis [30]. Given the non-linear nature of the quality space, a feature contributes both individually and through its interactions with other features. Therefore, permutation analysis correctly captures the importance addressing both these contributions. The approach systematically shuffles a given feature and explores the statistical impact of such manipulation on AQM outcomes. If a feature does not contribute significantly to the quality metric computation, there would be no incremental difference in the error from true label before and after shuffling a particular feature across the entire test set. Moreover, the sample distributions of quality metric scores would not be significantly different. Table 1 reports the tabulation of mean and standard errors along with statistical significance of disruption by shuffling each of the proposed features in the model. A ratio mean greater than 1 ensures the incremental nature of error when shuffling a certain feature and hence validates its importance. Moreover, to confirm this 'positive contribution' per feature is significant, the p-values reported in Table 1 are the results of paired t-test on AQM samples before and after shuffling a certain feature.



a) Addtive Linear Distortion b) Non-Linear Distortion

Fig. 2. Sensitivity of the obtained quality metric with the signal to noise ratio (SNR) across a plethora of noises (both linear and non-linear).

5. Sensitivity to ambient noise

5.1. Additive distortions

Signal to Noise Ratio is the most commonly used metric to quantify noise content in a signal. While this estimate requires information about the energy of the noise signal, this is not readily available in auscultation recordings. An evaluation of the relationship between AQM and signal to noise ratio can be done with controlled contamination where we systematically corrupt high quality lung sounds with varying degrees of noise levels, using both stationary and non-stationary profiles. We consider white noise as well as plausible noises in a clinic or ER setting (chatter and ambient noises [25]). Fig. 2(a) shows the correspondence between AQM and SNR levels for all 3 noise types. All 3 conditions (ambient noise, speech, and white noise) reveal a high linear correlation of 0.787, 0.9289, and 0.9439 (p-values: $5e^{-4}$, $< e^{-10}$, and $2e^{-4}$) respectively. The high non-stationary nature of ambient noise backgrounds creates highly variable signal profiles which vary the masking levels in any given segment hence allowing the auscultation to 'peak through' some moments of the signal leading to a less linear correspondence between AQM and SNR values, which are both averaged across many segments of 2 s-long signals.

5.2. Nonlinear distortions

The analysis also explores nonlinear clipping which is common during lung sound collection, arising from excessive friction of the transducer/diaphragm against the chest or clothing. The distortion is further exacerbated with uncooperative patients, particularly in pediatric settings. Due to its very transient nature, clipping shares a lot of spectrotemporal characteristics of crackle sounds [31], so false classification of a normal as adventitious is more likely with a prominent presence of clipping distortions. We evaluate the sensitivity of AQM with increasing degrees of clipping, by gradually saturating the signal envelope energy at different percentiles. Fig. 2(b) reveals a drastic drop of AQM values following an initial clipping, then a gradual drop afterward. It is worth noting that even at dramatic levels of clipping (close to 0), zero crossings in the signal are still preserved hence allowing some level of signal information to be maintained, though at low AQM values of 0.2. The estimated AQM values across these different ratios of clipping also exhibit a high linear correlation of 0.9570 (p-value $< e^{-10}$).

5.3. Inherent patient distortions

A third factor highly affecting auscultation quality is the state of the patient, where agitation and crying (particularly in pediatric patients) can affect signal integrity. Auscultations in presence of crying are particularly challenging because signal profiles reveal music sound characteristics that are easily confusable with wheezing sounds. Fig. 3 highlights the difficulties with this type of interference and shows an

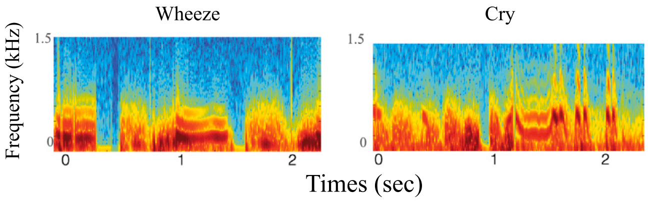


Fig. 3. Example of Wheeze and Cry spectrograms to visualize the similarity in spectral cues between an abnormal lung sound and an inherent patient distortion that is not disease indicative.

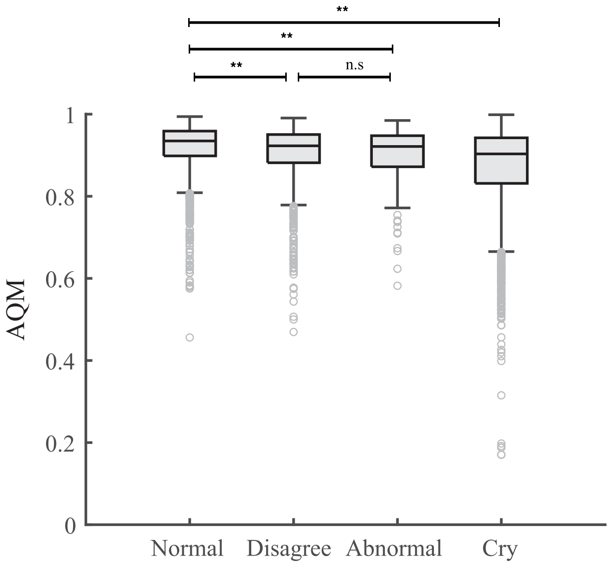


Fig. 4. Comparison of the performance of AQM across a multitude of clinical annotations: agreement, abnormality, and the presence/absence of cries.

example of a wheezing segment on the left and a crying segment visually depicting a similar spectral profile over time. Quantitatively, these two examples reflect acoustic properties that make them hard to distinguish. For instance, the empirical percentile (w.r.t the training feature sample space used to obtain AQM) for Bandwidth is 9% for the wheeze segment and 11% for the crying segment. Similarly, Pitch values are highly confusable (Wheeze: 22% & Cry: 28%) showing that both segments share similar spectro-temporal traits that make them lie closer to the empirical acoustic distributions used to train the AQM network. In contrast, the embedding analysis provides a counter-point to highlight differences between these segments. In the example shown in Fig. 3, the embedding ‘Mean Error’ reveals a closer match of the wheezing segment (21%) vs a clear distinction with the crying segment (72%) relative to training sample distributions. In order to quantify the effect of crying segments in the entire dataset, we use annotations from the expert listening panel identifying crying segments (see Methods in Section 2), and evaluate the distribution of AQM values for cry segments. Fig. 4-fourth box reports the variability of signal quality of auscultation segments contaminated by severe crying, with a mean of 0.87 and a variance of 0.011. Comparing the signal quality of crying segments against normal high-quality auscultations shows a statistically significant difference (p-value: $2.92e^{-69}$ using the Wilcoxon rank sum test).

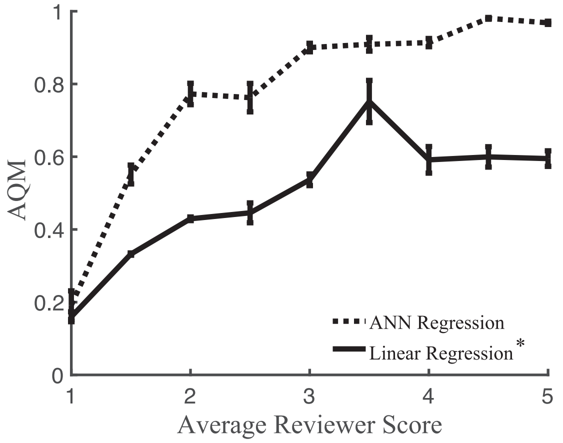


Fig. 5. Evaluation of proposed nonlinear quality metric against Expert Clinical Opinion (dashed line). AQM obtained by linear regression on a narrower feature set analyzed in [21] is reported as a solid line.

6. Clinical evaluation

6.1. AQM clinical validation

To further validate the AQM predictions, we compare AQM scores against the judgments of a panel of expert listeners on a quality scale of 1 to 5. Fig. 5 reports the average score of expert opinions versus AQM values on the same auscultation signals; and yields a high Pearson correlation of 0.831 (p = $1.39e^{-30}$). This correlation improves on a published linear quality metric [21] which yields a correlation of 0.76 ($p < 10^{-4}$) further supporting the need for a nonlinear mapping of acoustic and embedding features into a quality space.

6.2. Quality of normal and abnormal auscultations

Fig. 4 reports the quality measures of the high quality data by contrasting the average scores of normal versus abnormal lung sounds, as evaluated by the panel of expert listeners. Both classes yield average AQM values above 0.9 though there is a statistically significant difference between normals (mean 0.92) and abnormals (0.902) (Wilcoxon rank sum test, $3.6e^{-4}$). The slight drop in AQM values for abnormals is expected given that they share spectral profiles of noise signals, though the AQM also relies on high-level embeddings which maintain high scores for both groups of signals.

6.3. Clinical disagreement

We also analyze the relationship between auscultation quality and the clinical disagreement between expert reviewers. In many instances, reviewers disagreed on whether a signal is normal or contained adventitious segments. The AQM score population mean of the recordings with disagreement is 0.9045 and is significantly lower (Wilcoxon Rank Sum test p-value: $3.63e^{-12}$) than the normal population with clinical consensus (Fig. 4). Comparing the quality of disagreed recordings with abnormal recordings, the aggregate quality of abnormals is slightly lower than the disagreed recordings (the difference in population means is 0.0026) and this difference is not significant. The presence of ‘abnormal-like’ patterns might explain the disagreement in the diagnosis in the first place. However, the AQM still maintains a high value of > 0.9 confirming that the quality metric is primarily driven by signal quality rather than clinical evaluation of the signals.

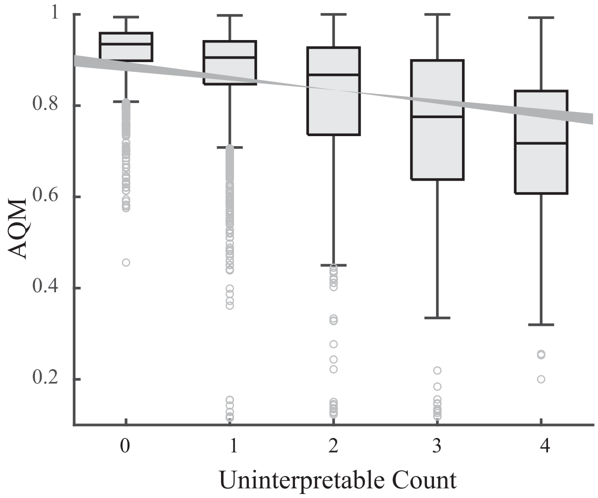


Fig. 6. Analysis of variation in AQM score with a decrease in interpretability measured by the number of annotators that have marked a recording clinically uninterpretable.

6.4. Non-interpretable auscultations

The presence of various forms of distortions and masking can render an auscultation signal clinically uninterpretable. While this notion of clinical interpretability is not binary, it partially reflects the integrity of the auscultation signal in addition to the inherent variability in human hearing and individual training of medical experts. We analyze the drop in AQM values with degree of uninterpretability defined by the number of expert reviewers marking a lung segment not clinically interpretable. A signal where 4 reviewers agree as not clinically viable is deemed more uninterpretable as compared to a signal where only one reviewer expressed dissatisfaction with the signal quality. Fig. 6 reveals a gradual drop in signal quality with an increased level of uninterpretability. A bootstrapping analysis to evaluate the inclination slope across 50 random sub-samplings of pools of 100 recordings reveals a statistically-significant negative drop -0.2208 ± 0.05.

7. Temporal sensitivity

One of the main uses of an auscultation quality measure is the ability to flag a segment on the fly (while being recorded) and urge the clinician to re-acquire the signal under better circumstances (calming down the patient, moving to a quieter room, stabilizing the stethoscope better). Given an analysis segment of 2 s with 50% overlap, we consider a sudden onset of noise and evaluate whether the AQM shows a concomitant drop. Fig. 7-inset shows a noise profile at −5 dB introduced in a high quality signal. We evaluate the 'response time' of the AQM over a wide range of such signals with noise introduced at $t_0 = 0$ s.

For this experiment, we corrupt signals with noises from BBC Database and white noise at a −5 dB signal-to-noise ratio. A random onset after the first two and before the last four seconds of the signals is chosen for this corruption. Ensuring the first two seconds always have 'high-quality' helps capture the fall in AQM scores by starting at an optimal value. Since the window is of 2-seconds, having noise contamination at least twice its size would confirm the complete overlap of the moving window and noise contaminated signal achieving the minimum possible AQM. This is validated by the statistical insignificance of the minimum AQM of signals corrupted with noise durations above 4 seconds and of signals completely corrupted. In Fig. 7, we notice the average trend of AQM moving score from two-seconds before the onset of noise. An average 60% drop in the AQM score is observed within 2.3 seconds of onset.

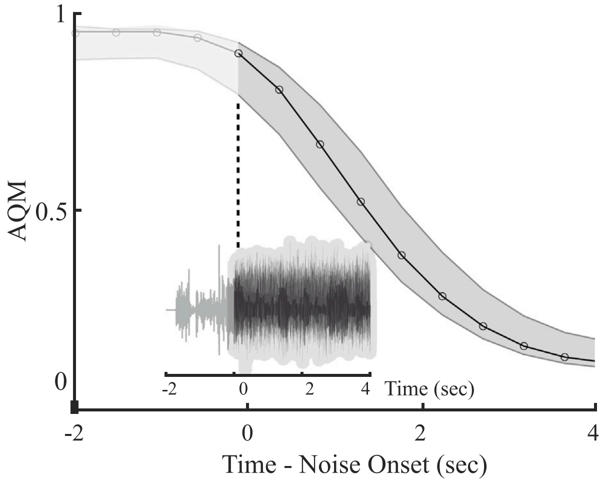


Fig. 7. Aggregate analysis of the sensitivity of the obtained AQM to the onset of corruption in recordings. An example of an audio waveform with such corruption from the onset is presented in the inset.

8. Discussion

The need for an auscultation quality metric stems from the presence of extreme variability in the data collection of auscultation signals and a lack of standardization. Although the denoising algorithms try to suppress the environmental noise, extreme noises like the subject's cry, reverberation, and electronic stethoscope sounds in a busy clinical environment are not completely eliminated [16]. It is to be noted that several other metrics like Segmental Signal to Noise Ratio, Normalized Covariance Metric, and Coherence Speech Intelligibility Index II like in the case of [32] evaluate the quality of lung sounds. But all these metrics would need access to the noise channel which is an impractical expectation to have for the recorded auscultation signals. The designed quality metric is working on par with these noise-dependent criteria in our analysis without actually looking at the original signal on our artificially corrupted dataset.

Which aspects of auscultation are important for quality judgment? Eight spectrotemporal features are chosen to best capture the quality profile of a recording in conjunction with its spectral properties. In [33], the authors observed a much better classification for normal and abnormal lung sounds on features that capture the rate-scale cortical representation of the recording. To capture this comprehensive profile of differences in normal and abnormal lung sounds, we look at both the rate of change of frequency content and how narrowband or broadband (scale) the spectral content is. Therefore, along with the entire spectral energy of the recordings which tends to be higher in the presence of noise, we also consider the energy content at different scales and rates. Pitch is considered as one of the features because it best captures background chatter, a common environmental noise. Prolonged abnormal lung sounds with a duration of more than 250 ms (wheezing) fall under narrowband processes while short explosive sounds (crackles) were found to be broadband signals. Bandwidth is included to account for this vast range of clean abnormal signals. Spectral Flatness is the ratio of a geometric mean and an arithmetic mean. Since geometric mean is at least as large as the arithmetic mean, a maximum Spectral Flatness measure of one is obtained only when each frequency channel has the exact same response over time (ex: white noise). Auscultation signals by themselves have a frequency range of 100 to 1000 Hz with a dip at 600 Hz. Considering the cut-off frequency of recordings is 4000 Hz, their profile is in no way uniform. In the presence of high-frequency noises, the frequency content gets flatter increasing the computed spectral flatness. In the same vein, we design data-driven features to encapsulate the contrast with low-quality auscultations from high quality sounds (both normal and abnormal). For an autoencoder capturing the profile of high-quality data, we expect a higher reconstruction error (ϵ) when a poor quality signal is sent. Similarly, given the compact nature of feature space, the mean feature profile of high quality signal and a poor quality signal feature would be farther apart and this distance is reflected by Mean Feature Error (μ).

8.1. Conclusion

In this work, we improve upon the features looked at to give a better assessment of auscultation quality and more importantly, obtain the auscultation metric in a non-linear fashion. This is evaluated by analyzing the Quality Metric on unseen non-additive noises like clipping in the artificially contaminated analysis. We further look at the trends of the obtained metric on clinical observations of expert annotators indicating the degree of intelligibility of the recordings or the presence of cries. Since abnormal lung sounds share similar spectrotemporal properties as the ambient noise, a slight decrease in their auscultation quality scores is observed compared to those of the normal recordings. But the overall scores are still high indicating a high quality signal. Similar to the former analysis, when looking at the degree of agreement amongst the reviewers on the final assessment of the type of lung sound, recordings with disagreement have slightly lower scores compared to the highly agreed upon data but higher than those of agreed upon abnormal lung sounds indicating adventitious like activity paved way to a slightly lower AQM. However, disagreement on data does not necessarily imply poorer quality which is validated by the overall higher quality metric. We delve into the direct correlation between low-quality scores and high clinical uninterpretability. The temporal sensitivity of the developed AQM is reviewed so that it can be used as a real-time indicator for the data collector to recollect a signal, in the case of integrity corruption. This no-reference nonlinear auscultation specific quality metric is robust to a plethora of environmental noises, complies with the clinical adjudication, and is sensitive to temporal corruptions ensuring its deployment potential to act as an entity to oversee auscultation collection.

CRediT authorship contribution statement

Annapurna Kala: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing. **Eric D. McCollum:** Methodology, Validation, Formal analysis, Data curation, Writing. **Mounya Elhilali:** Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

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. We would also like to thank Dr. Amyna Hussein for her assistance in the clinical evaluation of the auscultation signals. We also acknowledge support from NIH, United States 1R01HL163439, NIH, United States U01AG058532 and ONR, United States N00014-23-1-2050.

Algorithm 1 Algorithm for Spectral Features

```
for $w = 1 : W$ do
    for $t = 1 : T$ do
        Initialize Spectral Aggregates
        $ArithmeticMean = 0$
        $GeometricMean = 1$
        $SquaredAggregate = 0$
        $SquaredDiff = 0$
        $Brightness = 0$
        for $f = 1 : F$ do
            Compute Spectral Aggregates
            $ArithmeticMean \mathrel{+}= \frac{S_w(t,f)}{F}$
            $GeometricMean \mathrel{*}= S_w(t,f)^{\frac{1}{F}}$
            $SquaredAggregate \mathrel{+}= S_w(t.f)^2$
            $SquaredDiff \mathrel{+}= (S_w(t,f+1) - S_w(t,f))^2$
            $Brightness \mathrel{+}= f * S_w(t.f)^2$
        end for
        for $f = 1 : F$ do
            Compute Spectro-temporal Features
            $\boldsymbol{E[S]} \mathrel{+}= \frac{1}{W} S_w(t,f)$
            $\boldsymbol{E[\hat{S}]} \mathrel{+}= \frac{1}{W} \sum_{\Omega=0.25}^{8} S_w(t,f) * h(f,\Omega)$
            $\boldsymbol{E[HR]} \mathrel{+}= \frac{1}{W} \sum_{\omega=30} S_w(t,f) * g(t,\omega)$
            $\boldsymbol{E[LR]} \mathrel{+}= \frac{1}{W} \sum_{\omega=1}^{30} S_w(t,f) * g(t,\omega)$
            $\boldsymbol{BW} \mathrel{+}= \frac{1}{W} \frac{|f - \frac{Brightness}{SquaredAggregate}| * S_w(t,f)}{F * ArithmeticMean}$
        end for
        $\boldsymbol{SF} \mathrel{+}= \frac{1}{W} \frac{GeometricMean}{ArithmeticMean}$
        $\boldsymbol{SI} \mathrel{+}= \frac{1}{W} \frac{SquaredDifference}{SquaredAggregate}$
    end for
    $\hat{F_o} \mathrel{+}= \frac{1}{W} argmax_{P_k} S_w.T_k$
end for
```

Appendix. Acoustic feature extraction algorithm

See Algorithm 1.

References

[1] P.J. Bishop, Evolution of the stethoscope, J. R. Soc. Med. 73 (6) (1980) 448–456.

[2] M. Elhilali, J.E. West, The stethoscope gets smart, IEEE Spectr. 56 (02) (2019) 36–41, http://dx.doi.org/10.1109/MSPEC.2019.8635815.

[3] R.X.A. Pramono, S. Bowyer, E. Rodriguez-Villegas, Automatic adventitious respiratory sound analysis: A systematic review, PLoS ONE 12 (5) (2017) 1–43, http://dx.doi.org/10.1371/journal.pone.0177926.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

[5] G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 20 (1) (2012) 30–42, http://dx.doi.org/10.1109/TASL.2011.2134090.

[6] P. Pujol, S. Pol, C. Nadeu, A. Hagen, H. Bourlard, Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system, IEEE Trans. Speech Audio Process. 13 (2005) 14–22.

[7] M. Aykanat, Ö. Kılı ç, B. Kurt, S. Saryal, Classification of lung sounds using convolutional neural networks, EURASIP J. Image Video Process. 2017 (1) (2017) http://dx.doi.org/10.1186/s13640-017-0213-2.

[8] D. Perna, A. Tagarelli, Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2019-June, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 50–55, http://dx.doi.org/10.1109/CBMS.2019.00020.

[9] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, A. Shalyto, Noise masking recurrent neural network for respiratory sound classification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11141 LNCS, Springer, Cham, 2018, pp. 208–217, http://dx.doi.org/10.1007/978-3-030-01424-7_21.

[10] J. Acharya, A. Basu, Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning, IEEE Trans. Biomed. Circuits Syst. 14 (3) (2020) 535–544, http://dx.doi.org/10.1109/TBCAS.2020.2981172.

[11] B.M. Rocha, D. Pessoa, A. Marques, P. Carvalho, R.P. Paiva, Automatic classification of adventitious respiratory sounds: A (un)solved problem? Sensors 21 (1) (2020) 57, http://dx.doi.org/10.3390/s21010057.

[12] P. Gomes, S. Frade, A. Castro, R. Cruz -Correia, M. Coimbra, A proposal to incorporate digital auscultation and its processing into an existing electronic health record, in: HEALTHINF 2015 - 8th International Conference on Health Informatics, Proceedings; Part of 8th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2015, SciTePress, 2015, pp. 143–150, http://dx.doi.org/10.5220/0005222901430150.

[13] N.Q. Al-Naggar, A new method of lung sounds filtering using modulated least mean square—Adaptive noise cancellation, J. Biomed. Sci. Eng. 6 (2013) 869–876.

[14] K.K. Guntupalli, P.M. Alapat, V.D. Bandi, I. Kushnir, Validation of automatic wheeze detection in patients with obstructed airways and in healthy subjects, J. Asthma 45 (10) (2008) 903–907, http://dx.doi.org/10.1080/02770900802386008.

[15] J. Li, Y. Hong, Wheeze detection algorithm based on spectrogram analysis, in: 2015 8th International Symposium on Computational Intelligence and Design, ISCID, 1, 2015, pp. 318–322, http://dx.doi.org/10.1109/ISCID.2015.310.

[16] D. Emmanouilidou, M. Elhilali, Characterization of noise contaminations in lung sound recordings, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2013, pp. 2551–2554, http://dx.doi.org/10.1109/EMBC.2013.6610060.

[17] M.J. Mussell, The need for standards in recording and analysing respiratory sounds, Med. Biol. Eng. Comput. 30 (2) (1992) 129–139, http://dx.doi.org/10.1007/BF02446121, URL https://pubmed.ncbi.nlm.nih.gov/1453777/.

[18] D.B. Springer, T. Brennan, N. Ntusi, H.Y. Abdelrahman, L.J. Zühlke, B.M. Mayosi, L. Tarassenko, G.D. Clifford, Automated signal quality assessment of mobile phone-recorded heart sound signals, J. Med. Eng. Technol. 40 (7–8) (2016) 342–355, http://dx.doi.org/10.1080/03091902.2016.1213902.

[19] C.F. Camm, N. Sunderland, A.J. Camm, A quality assessment of cardiac auscultation material on YouTube, Clin. Cardiol. 36 (2) (2013) 77–81, http://dx.doi.org/10.1002/clc.22080.

[20] E. Grooby, J. He, J. Kiewsky, D. Fattahi, L. Zhou, A. King, A. Ramanathan, A. Malhotra, G.A. Dumont, F. Marzbanrad, Neonatal heart and lung sound quality assessment for robust heart and breathing rate estimation for telehealth applications, IEEE J. Biomed. Health Inf. PP (2020) http://dx.doi.org/10.1109/JBHI.2020.3047602.

[21] A. Kala, A. Husain, E.D. McCollum, M. Elhilali, An objective measure of signal quality for pediatric lung auscultations, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2020, pp. 772–775, http://dx.doi.org/10.1109/EMBC44109.2020.9176539.

[22] E. McCollum, D. Park, N. Watson, C. Focht, C. Bunthi, B. Ebruke, M. Elhilali, D. Emmnouilidou, L. Hossain, D. Moore, A. Mudaua, J. Mulindwa, J. West, K. O'Brien, D. Feikin, L. Hammitt, Digitally-Recorded Lung Sounds and Mortality Among Children 1-59 Months Old with Pneumonia in the Pneumonia Etiology Research for Child Health Study, 2017.

[23] D. Emmanouilidou, E.D. McCollum, D.E. Park, M. Elhilali, Computerized lung sound screening for pediatric auscultation in noisy field environments, IEEE Trans. Biomed. Eng. 65 (7) (2018) 1564–1574, http://dx.doi.org/10.1109/TBME.2017.2717280.

[24] E.D. McCollum, D.E. Park, N.L. Watson, W.C. Buck, C. Bunthi, A. Devendra, B.E. Ebruke, M. Elhilali, D. Emmanouilidou, A.J. Garcia-Prats, L. Githinji, L. Hossain, S.A. Madhi, D.P. Moore, J. Mulindwa, D. Olson, J.O. Awori, W.P. Vandepitte, C. Verwey, J.E. West, M.D. Knoll, K.L. O'Brien, D.R. Feikin, L.L. Hammit, Listening panel agreement and characteristics of lung sounds digitally recorded from children aged 1–59 months enrolled in the pneumonia etiology research for child health (PERCH) case–control study, BMJ Open Respir. Res. 4 (1) (2017) e000193, http://dx.doi.org/10.1136/bmjresp-2017-000193.

[25] BBC, The BBC sound effects library, 1990.

[26] T. Chi, P. Ru, S.A. Shamma, Multiresolution spectrotemporal analysis of complex sounds, J. Acoust. Soc. Am. 118 (2) (2005) 887–906.

[27] N. Huang, M. Elhilali, Auditory salience using natural soundscapes, J. Acoust. Soc. Am. 141 (3) (2017) 2163–2176, http://dx.doi.org/10.1121/1.4979055.

[28] S.A. Shamma, D.J. Klein, The case of the missing pitch templates: How harmonic templates emerge in the early auditory system, J. Acoust. Soc. Am. 107 (5) (2000) 2631–2644.

[29] A.H. Gray, J.D. Markel, A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis, IEEE Trans. Acoust. Speech Signal Process. 22 (3) (1974) 207–217, http://dx.doi.org/10.1109/TASSP.1974.1162572.

[30] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: A corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347, http://dx.doi.org/10.1093/BIOINFORMATICS/BTQ134.

[31] Z. Moussavi, Respiratory sound analysis [introduction for the special issue] IEEE Eng. Med. Biol. Mag. 26 (1) (2007) 15, http://dx.doi.org/10.1109/MEMB.2007.289116.

[32] I. McLane, D. Emmanouilidou, J.E. West, M. Elhilali, Design and comparative performance of a robust lung auscultation system for noisy clinical settings, IEEE J. Biomed. Health Inf. 25 (7) (2021) 2583–2594, http://dx.doi.org/10.1109/JBHI.2021.3056916.

[33] D. Emmanouilidou, M. Elhilali, Rich representation spaces: Benefits in digital auscultation signal analysis, in: 2016 IEEE International Workshop on Signal Processing Systems, SiPS, IEEE, 2016, pp. 69–73, http://dx.doi.org/10.1109/SiPS.2016.20.