Automatic Croup Diagnosis Using Cough Sound Recognition

Roneel V. Sharan, *Member, IEEE*, Udantha R. Abeyratne*, *Senior Member, IEEE*,
Vinayak R. Swarnkar, and Paul Porter

Abstract—Objective: **Croup, a respiratory tract infection common in children, causes an inflammation of the upper airway restricting normal breathing and producing cough sounds typically described as seal-like "barking cough". Physicians use the existence of barking cough as the defining characteristic of croup. This work aims to develop automated cough sound analysis methods to objectively diagnose croup.** *Methods:* **In automating croup diagnosis, we propose the use of mathematical features inspired by the human auditory system. In particular, we utilize the cochleagram for feature extraction, a time-frequency representation where the frequency components are based on the frequency selectivity property of the human cochlea. Speech and cough share some similarities in the generation process and physiological wetware used. As such, we also propose the use of mel-frequency cepstral coefficients which has been shown to capture the relevant aspects of the short-term power spectrum of speech signals. Feature combination and backward sequential feature selection are also experimented with. Experimentation is performed on cough sound recordings from patients presenting various clinically diagnosed respiratory tract infections divided into croup and non-croup. The dataset is divided into training and test sets of 364 and 115 patients, respectively, with automatically segmented cough sound segments.** *Results:* **Croup and non-croup patient classification on the test dataset with the proposed methods achieve a sensitivity and specificity of 92.31% and 85.29%, respectively.** *Conclusion:* **Experimental results show significant improvement in automatic croup diagnosis against earlier methods.** *Significance:* **This work has the potential to automate croup diagnosis based solely on cough sound analysis.**

*Index Terms—***Cough sound recognition, croup, mel-frequency cepstral coefficients, sequential feature selection, support vector machines, time-frequency image.**

I. INTRODUCTION

CROUP is a viral infection of the respiratory tract which obstructs the upper airway and can be life-threatening when severe [1]. As summarized in [2], it is the most common form of airway obstruction in children between 6 months and 6 years of age, peaking between 1 and 2 years of age. A two-year Australian study (2008–2009) in children aged 0–14 years encountered croup in 1.2% of subjects or about 154,000 times per year [3]. It was determined to be most prevalent in children aged 1–4 years.

The inflammation of the upper airway caused by the infection restricts normal breathing which results in a 'croupy' or 'barking' cough sometimes accompanied by stridor, hoarse voice, and respiratory distress [4, 5]. The additional clinical signs are, however, not unique to croup and may be present in other obstructive respiratory diseases as well. Therefore, the distinctive cough is the primary clinical feature used in clinical practice to diagnose croup. Physicians make a subjective judgment on the 'croupiness' or 'barkingness' of the cough after listening to it. Croup diagnosis is, therefore, limited to human perception and dependent on the skills of the clinician.

In this paper, we aim to remove this subjectivity in croup diagnosis. We propose to automate croup detection through objective analysis of cough sounds using signal processing and machine learning techniques.

The potential benefits of this approach are manifold. One of the main advantages is that lay people with minimal or no clinical training will be able to use the technology. The technology can be implemented on a smart phone platform such as the iPhone without a need for a network connection and deployed in remote regions of the world. It will also have the potential for use in urban centers of the developed world in applications such as triaging in large health facilities, disaster medicine, and in screening in airports. Croup affects children and the symptoms of the disease often appear at night [5, 6]. The sudden onset of symptoms causes many parents to panic and immediately visit an emergency department [6]. The automated diagnostic tool could be used to better manage such situations through a tele-consultation approach.

In this paper, we propose the use of mathematical features inspired by the human auditory system. The conventional time-frequency plot, spectrogram, has evenly spaced frequency bins with constant bandwidth. This is not appropriate for modeling the cochlea of the human auditory

system, which uses a substantially different philosophy in decomposing a sound signal to its frequency components prior to further analysis. The cochlea uses more frequency bins in the lower frequency range with narrow bandwidth and less frequency components in the higher frequency range with wider bandwidth. In this paper, we achieve this by using a time-frequency representation where the frequency components are determined using a bio-inspired gammatone filter. The gammatone filter models the frequency selectivity of the human cochlea. The resulting time-frequency image is referred to as a cochleagram and the resulting feature, which captures the statistical distribution, as the cochleagram image feature (CIF) [7].

This CIF enables us to capture low-frequency information with a higher granularity, without sacrificing information available in higher frequency ranges. The 'barkingness' of a croup cough is expected to add distinguishing features at lower frequencies, helping with the sensitivity of diagnosis while high frequency information should improve the specificity of diagnosis. Consumer audio recording devices of today are able to capture audio data covering the entire range of the human auditory spectrum.

Continuing with our philosophy of drawing inspiration from the human auditory system, we propose augmenting the CIF feature with Mel-frequency Cepstral Coefficients (MFCC) [9]. MFCCs capture the perceptual properties of the auditory system and have found wide acceptance as a powerful feature in automatic speech recognition (ASR) systems.

In this paper, we propose the use of linear MFCC in preference to the conventional log MFCC. MFCC utilize mel-filter banks and unlike log MFCC, linear MFCC does not apply any compression to the filter bank energies. The peak of the filter bank energies plays a key role in characterizing a sound signal. However, the conventional log compression can produce high variations in the output for low energy components [8]. This led to the introduction of root compressed MFCC [9], a technique shown to be more robust in ASR. Linear MFCC is a special case of root compressed MFCC where the filter bank energies are raised to the power of one or, essentially, left uncompressed.

We experiment with feature combination as a way of increasing the diagnostic performance of our technology. Feature combination is a common technique in trying to improve the classification performance. Feature addition, particularly to MFCCs, has been shown to be effective in resolving various cough sound classification problems. In our earlier works [10, 11] we illustrated the usefulness of MFCCs (in combination with other features) in diagnosing diseases such as pneumonia. A similar combined approach has also been taken in diagnosing pertussis [12]. For the problem considered in this work, we propose the combination of linear MFCC and CIF.

The performance of the technology can be improved by a careful selection of features and a good classifier. Our choice of classifiers is based on their simplicity of implementation on a smartphone and the diagnostic performance. Logistic regression model (LRM) is a well-known linear classifier which has been recently used in cough analysis [10-12]. In [10, 11], we found the training and testing process for LRM to be very simple and fast and is, therefore, our baseline classifier for this work. In addition, we consider support vector machines (SVM). SVM has particularly been shown to be effective on small datasets. While nonlinear SVM requires tuning of classifier parameters during training phase, the testing phase is much simpler and quicker.

Feature selection targeting the reduction of the dimension of the feature-vector could present many benefits such as improved classifier generalization performance, reduced evaluation time, and less storage space. In this work, we explored two feature reduction methods to optimize the performance of the algorithm.

In the first approach, we used p-value statistics for each individual feature as available in training the LRM classifier. This is inspired by our success [10] with this approach in the cough-based diagnosis of pneumonia. In the second approach, we use the first approach as our baseline feature selection method and augment it with a backward sequential feature selection, a technique that is specifically targeted at improving the classification performance. Backward sequential feature selection starts with all the features in the model and systematically removes one feature at a time until no further improvement can be achieved in the classification performance. We also extend this feature selection approach to the design of the SVM classifier.

The fully automated croup diagnosis system we envision requires an automated cough segmentation module as its front-end. While we used only manually segmented cough sound signals in our previous works [10, 11], in this paper we also test our methods with automatically segmented coughs. The automatic cough segmentation algorithm we developed is described in Appendix A. The automatic segmentation algorithm is intended as a proof of concept that a fully automated clinical system could be developed. However, the major focus of the current paper is on the post-segmentation processing to diagnose croup.

The performance of the proposed methods is evaluated on a clinical database of cough sounds recorded in real-world environments in hospitals in Australia. Our database consists of two data sets; Dataset A (364 patients) and Dataset B (115 patients). Dataset A is used to train and cross-validate our models whereas Dataset B is used solely as an independent prospective set to test the models developed on Dataset A. The two data sets are mutually exclusive both at the patient and individual cough levels.

The rest of this paper is organized as follows. Section II gives an overview of the feature extraction methods utilized in this work. Section III gives an overview of the classification and feature selection methods. A description of the cough sound database used in this work is given in Section IV and experimental evaluation is performed in section V. Conclusion and recommendations are given in section VI.

A preliminary version of this work has been reported [13].

II. FEATURE EXTRACTION

This section describes the feature extraction methods for MFCC and CIF with reference to Fig. 1.

A. MFCC

In computing MFCC features, the cough signal is divided into frames and discrete Fourier transform (DFT) is applied to the windowed frames as

$$X(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{\frac{-2\pi i k n}{N}}, \quad k = 0, 1, ..., N-1 \quad (1)$$

where N is the length of the frame, $x(n)$ is the time domain signal, $w(n)$ is the window function, and $X(k)$ is the kth harmonic corresponding to the frequency $f(k) = kF_s/N$, F_s is the sampling frequency.

MFCCs utilize mel-filter banks, or triangular bandpass filters, which are equally spaced on the mel-scale [14]. The adjacent filters overlap such that the lower and upper ends of the mth filter are located at the centre frequency of the $m-1$ and $m+1$ filter, respectively.

The conventional log MFCCs are obtained as the discrete cosine transform (DCT) of the log compressed filter bank energies given as

$$c(i) = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} \log\big(E(m)\big) \cos\left(\frac{\pi i}{M}(m-0.5)\right) \quad (2)$$

where $i = 1, 2, ..., l$, l is the order of the cepstrum, $E(m)$ is the filter bank energies of the mth filter, and M is the total number of mel-filters.

Root cepstral coefficients [9] takes a similar computation method except that root compression is applied to the filter bank energies, $[E(m)]^\gamma$ where $0 < \gamma \le 1$, instead of log compression. Linear MFCC is a special case of root cepstral coefficients where $\gamma = 1$, that is, no compression is applied to the filter bank energies.

B. CIF

In this time-frequency representation, the signal is broken into different frequencies which are those recognized by the cochlea and hair cells. This frequency selectivity is modeled by the gammatone filter which is a series of bandpass filters with impulse response [15]

$$h(t) = At^{j-1}e^{-2\pi Bt}\cos(2\pi f_c t + \phi) \quad (3)$$

where A is the amplitude, j is the order of the filter, B is the bandwidth of the filter, f_c is the center frequency of the filter, ϕ is the phase, and t is the time.

The equivalent rectangular bandwidth (ERB) is used to describe the bandwidth of each cochlea filter in [15] which is given as
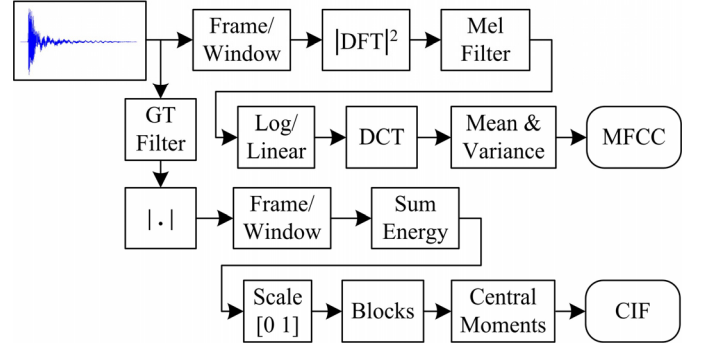

Fig. 1. An overview of computing MFCC and CIF.

$$f_{c,ERB} = \left[\left(\frac{f_{c,Hz}}{Q_{ear}}\right)^p + \big(B_{min}\big)^p\right]^{1/p} \quad (4)$$

where Q_{ear} is the asymptotic filter quality at high frequencies and B_{min} is the minimum bandwidth for low frequency channels. The bandwidth of a filter can then be approximated as $B = 1.019 \times f_{c,ERB}$. For this work, we only consider Greenwood's ERB model [16] which was shown to give the best classification performance in [7].

The human cochlea has thousands of hair cells which resonate at their characteristic frequency and at a certain bandwidth. The mapping between filter index and center frequency is determined as [17]

$$f_{cg} = -Q_{ear}B_{min} + \big(f_h + Q_{ear}B_{min}\big)e^{-gs/Q_{ear}} \quad (5)$$

where $g = 1, 2, ..., G$, G is the number of gammatone filters, f_h is the maximum frequency in the filter bank, and s is the step factor given as

$$s = \frac{Q_{ear}}{G}\log\left(\frac{f_h + Q_{ear}B_{min}}{f_l + Q_{ear}B_{min}}\right) \quad (6)$$

where f_l is the minimum frequency in the filter bank.

Similar to [7] we use a 4th order gammatone filter with four filter stages and each stage a 2nd order digital filter as given in [17]. The gammatone filter was implemented using the Auditory Toolbox for Matlab [18].

A representation similar to the conventional spectrogram image is obtained by smoothing the time series associated with each frequency channel in the gammatone filter and then adding the energy in the windowed signal for each frequency component. These form the intensity values which are then scaled in the range [0 1] for feature extraction. The time domain signal, spectrogram, and cochleagram of a normal and croup cough sound signal are given in Fig. 2. The dominant frequency component, centered around 400 Hz, is suppressed in the spectrogram image but revealed more clearly in the
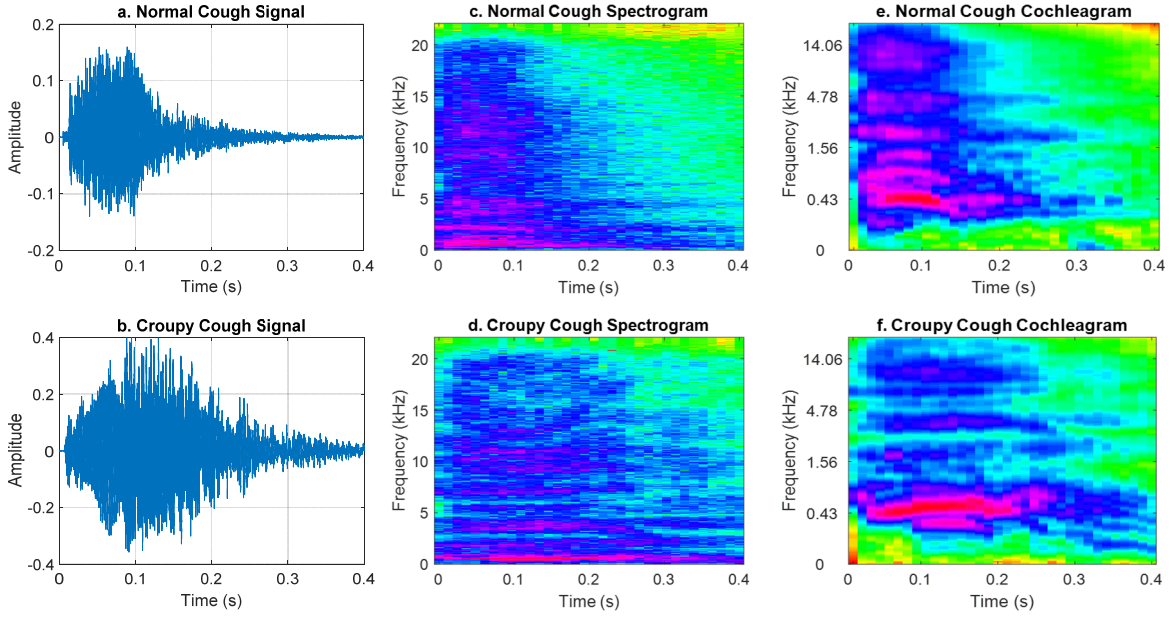
Fig. 2. (a) and (b) Time domain waveforms of a normal and a croupy cough, (c) and (d) their spectrograms, and (e) and (f): corresponding cochleagrams [13].
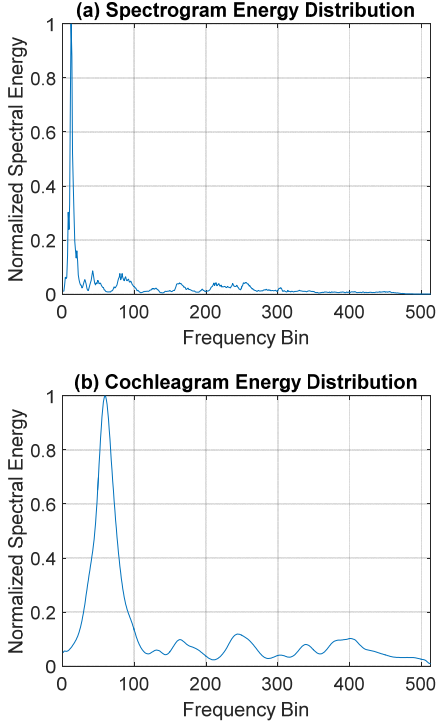

Fig. 3. Normalized spectral energy distribution of (a) spectrogram image and (b) cochleagram image.

cochleagram courtesy of more frequency components in the lower frequency range with narrow bandwidth. The frequency range in both the representations is 0 to 22,050 Hz, which is the Nyquist frequency.

The spectral energy distributions over the frequency bins for the spectrogram and cochleagram time-frequency representations are shown in Fig. 3(a) and (b), respectively. The spectrogram image has most of the spectral energy concentrated in the lower frequency bins, particularly between 0 and 20 with the other frequency bins generally carrying

significantly less energy. The cochleagram image has a different distribution with a wider energy spread making it more useful for feature extraction.

III. CLASSIFICATION AND FEATURE SELECTION

A. Logistic Regression Model (LRM)

LRM is a regression model where the dependent variable is categorical, the probability of which is estimated using one or more independent variables or features. The dependent variables in this work are croup ($Y = 1$) and non-croup ($Y = 0$). For a given feature vector $\mathbf{F} = [f_1\ f_2\ ...\ f_f]$, the probability that the output is croup ($Y = 1$) can be estimated using the logistic function given as

$$P(Y = 1 | \mathbf{F}) = \frac{e^v}{e^v + 1} \tag{7}$$

where

$$v = \beta_0 + \beta_1 f_1 + ... \beta_f f_f \tag{8}$$

and $\beta_0, \beta_1, ..., \beta_f$ are the regression coefficients.

B. Support Vector Machine (SVM)

SVM determines the optimal hyperplane to maximize the distance between any two given classes [19]. Given a set of S training samples belonging to two classes as $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_S, y_S)\}$, where $\mathbf{x}_s \in R^d$ is a d-dimensional feature vector representing the sth training sample, and $y_s \in \{-1, +1\}$ is the class label of $\mathbf{x}_S$. The optimal hyperplane can be determined by minimizing $\frac{1}{2}\|\mathbf{w}\|^2$ subject to $y_s(\mathbf{w} \cdot \mathbf{x}_s + b) \geq 1$, where $\mathbf{w} \in R^d$ is a normal vector to the hyperplane and b is a constant. The optimization is solved under the given constraints by the saddle point of the Lagrangian function.

For linearly nonseparable problems, the optimization can be generalized by introducing the concept of *soft margin* [19]. Nonlinear SVM is used in this work which maps the input vector $\mathbf{x}$ to a higher dimensional space $\mathbf{z}$ through some nonlinear mapping $\phi(\mathbf{x})$ chosen *a priori* to construct an optimal hyperplane. The *kernel trick* [20] is applied to create the nonlinear classifier where the dot product is replaced by a nonlinear kernel function $K(\mathbf{x}_s, \mathbf{x}_r)$ which computes the inner product of the vectors $\phi(\mathbf{x}_s)$ and $\phi(\mathbf{x}_r)$. A commonly used kernel function is Gaussian radial basis function (RBF), $K(\mathbf{x}_s, \mathbf{x}_r) = \exp(-\|\mathbf{x}_s - \mathbf{x}_r\|^2 / 2\sigma^2)$, where $\sigma > 0$ is the width of the Gaussian function.

The classifier for a given kernel function with the optimal separating hyperplane is then given as

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{s=1}^{S} \alpha_s y_s K(\mathbf{x}_s, \mathbf{x}) + b \right) \tag{9}$$

with α_s being the Lagrange multipliers.

C. Patient Classification

For the uth patient, the rth cough is classified as croup if its posterior probability is ≥ 0.5, that is,

$$C_u^r = \begin{cases} 1, & P \geq 0.5 \\ 0, & P < 0.5 \end{cases} \tag{10}$$

where $r = 1, 2, \ldots, R_u$, R_u is the total number of coughs for the uth patient, $u = 1, 2, \ldots, U$, U is the total number of patients.

The uth patient is then classified as having croup if one or more coughs are classified as croup, that is,

$$D_u = \begin{cases} 1, & \sum_{r=1}^{R_u} C_u^r \geq 1 \\ 0, & \sum_{r=1}^{R_u} C_u^r = 0 \end{cases} \tag{11}$$

In [10], cough index, defined as the fraction of coughs labeled as diseased, that is, croup, is used to classify a patient into disease and non-disease groups. While we used (11) to determine the patient disease classification in this work, we also experimented with cough index and a cough index of 0.1 was seen to give best results. The results were very similar to what was achieved using (11) which can be expected since the maximum number of coughs is restricted to 10 in this work.

D. Feature Selection

1) P-Value Statistics

With the LRM classifier, we use the *p*-value statistics to determine the significance of a feature dimension to the model, as inspired by the success of our earlier work in pneumonia diagnosis [10]. The *p*-value is computed for each feature with an output range [0 1], a low *p*-value indicating a higher significance and vice-versa. The average *p*-value is computed over the trained models in leave-one-out cross-validation. Different *p*-value thresholds were then applied to determine the model which produced the best classification performance.

2) Backward Sequential Feature Selection

In Step 1 of backward sequential feature selection, we use all the features to calculate the mean error rate using leave-one-out cross-validation. In Step 2, one feature dimension is removed at a time and the mean error rate calculated with the remaining features at each iteration. At the end of this step, the feature removal corresponding to the lowest mean error rate is removed. Step 2 is repeated with the remaining features. This process continues until no further improvement can be achieved in the error rate. The backward sequential feature selection process is terminated at this point and the remaining features are then utilized in training and testing the final models.

IV. COUGH SOUND DATABASE

A. Data Recording

Cough sounds were recorded from two clinical sites, Joondalup Health Campus (JHC) and Princess Margaret Hospital (PMH), both in Perth, Western Australia. Patient population has children suspected of respiratory illnesses such as pneumonia, asthma/RAD (Reactive Airway Disease), bronchiolitis, croup and upper respiratory tract infection (URTI). The human ethics committees of The University of Queensland, Joondalup Health Campus, and Princess Margaret Hospital had approved the study protocols and the patient recruitment procedure.

Patients fulfilling the inclusion criteria (presenting with cough, wheeze, shortness of breath, stridor, URTI and not satisfying the exclusion criteria (requiring respiratory support, no consent given) were recruited to the study. Healthy subjects, defined as children who did not have any symptom of respiratory disease at the time of measurement, were also recruited.

Cough sounds were recorded using an Apple iPhone 6s. Sound data were recorded at a sampling rate of 44,100 samples per second at a bit depth of 16-bits per sample. The smartphone recorder was placed approximately 50 cm away from the mouth of the patient and at an angle of approximately 45° (see Fig. 4).

Sound recordings were made in realistic clinical environments of these hospitals. Efforts were made to avoid procedurally preventable interferences such as background TV, loud background conversations, and adult coughs in recordings. However, the acquired audio files had interferences such as cries, footsteps, occasional speech and beeps from other medical instruments unavoidable in the clinical environment.

B. Database Overview

Our database consists of cough recordings and detailed clinical diagnostic information on each patient including the final diagnosis, clinical examination findings and laboratory as

Fig. 4. Recording cough sounds at the hospital using a smartphone (iPhone).

well as imaging outcomes. Demographic information was also available in a patient de-identified format.

The cough sound database used for the work of this paper is divided into two sets which are independent of each other and were recorded over different periods. Dataset A is our model training and validation dataset. It has a total of 364 pediatric patients belonging to two classes: croup (43 patients) and non-croup (321 patients). It is used for training and validating the models. Dataset B is our model testing dataset. It has a total of 115 pediatric patients belonging to two classes: croup (13 patients) and non-croup (102 patients). It is used for testing only, on the models developed on Dataset A.

Both datasets have multiple voluntary and spontaneous cough sound signals recorded by nurses. The cough sound signals for each patient are manually and automatically segmented with up to 10 coughs per patient included in the analysis. The procedure for auto segmentation of cough sound signals is described in Appendix A.

For both datasets, the class croup includes patients diagnosed with croup alone or croup comorbid with upper respiratory tract infection (URTI). The non-croup class includes patients with asthma/viral-induced wheeze, bronchiolitis, and pneumonia (atypical, bacterial, and viral). The non-croup class also includes URTI (as an isolated diagnosis and/or as a comorbidity of other non-croup diseases). All the respiratory tract infections used in the database have been diagnosed by clinicians at PMH and JHC using Australian clinical guidelines. Age, gender, and cough statistical data on croup and non-croup patients in Dataset A and Dataset B are given in Table I.

V. ALGORITHM TRAINING AND EVALUATION

In this section, we present an overview of our algorithm development and validation procedures followed by results obtained during model training and prospective testing.

A. Overview

Algorithm development and evaluation was carried out as described in Section II and Section III.

When extracting features from cough sounds, refer to (1), we used a frame size $N = 1024$ samples (23.22 ms) and Hamming window for $w(n)$. Frame-to-frame overlap of 50%

was used. We also experimented with dividing the cough sound signal into three equal parts, as in [10], and then following the same feature computation procedure. However, in the particular problem of diagnosing croup, we did not see any advantage in the classification performance over the sliding window method.

We used the leave-one-out cross validation technique to train and validate our models on the training data set (Dataset A). Once the model development is over, we fixed the parameters and trained the final model on the entire training dataset. Then we tested it on the prospective data set (Dataset B). Using the clinical diagnosis as the reference standard, we then calculated performance measures such as the sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and Cohen's kappa (κ). Except κ, all these values are reported as a percentage (%).

In the calculation of MFCC coefficients, we explored the effect of different number (M) of mel-filters in the range of 10-50. The best results were obtained at $M = 18$ for log MFCC and $M = 15$ for linear MFCC. As such, the feature vector for each frame is 54 dimensional for log MFCC, 18 cepstral coefficients plus the first and second derivatives [21], and, similarly, 45 dimensional for linear MFCC. The final feature vector is a concatenation of the mean and standard deviation values along each dimension resulting in a 108 dimensional final feature vector for log MFCC and 90 dimensional final feature vector for linear MFCC. We also experimented with cepstral mean and variance normalization and cepstral scaling before computing the mean and standard deviation for the final feature vector. Cepstral scaling was seen to improve the classification performance, particularly in the case of linear MFCC.

For the cochleagram image, to get the same image resolution along the frequency axis as the spectrogram image, that is $N/2$, the number of gammatone filters, G, is set to 512. The cochleagram image is divided into blocks and the second and third central moments are extracted as features in each block. These values are concatenated to form the final feature vector. Various number of blocks were experimented with and the best results obtained with 8×4 blocks, along the vertical and horizontal, respectively. This results in a 64 dimensional final feature vector.

We present results based on two different classifiers (LRM and SVM) as described in Section III. The LRM was chosen as a simple and effective linear classifier in cough analysis [10, 22], whereas the SVM approach (with a Guassian Radial basis Function (RBF) kernel) was used as a nonlinear approach suitable for the small datasets we had access to. SVM parameters, the penalty parameter and the width of the Gaussian function [19], were tuned using Bayesian optimization [23]. The aim in parameter tuning was to minimize the mean error of leave-one-out cross validation.

In this paper we obtain results with two different approaches in extracting coughs (cough segmentation) from the recorded audio streams; namely manual and automated cough picking.

TABLE I
STATISTICAL ANALYSIS OF AGE, GENDER, AND COUGH DATA OF CROUP AND NON-CROUP PATIENTS ON DATASET A AND DATASET B

		Dataset A			Dataset B		
		Croup	Non-Croup	Overall	Croup	Non-Croup	Overall
Age	< 12 months	6	40	46	3	22	25
	≥ 12 AND < 60 months	21	154	175	4	42	46
	≥ 60 AND < 120 months	11	108	119	4	27	31
	≥ 120 months	5	19	24	2	11	13
	Mean (months)	54.33 ± 42.53	55.50 ± 37.57	55.36 ± 38.20	62.31 ± 42.28	54.16 ± 43.15	55.08 ± 43.13
	Range (months)	5 – 148	0 – 192	0 – 192	8 – 161	0.5 – 177	0.5 – 177
Gender	Male	34	205	239	11	60	71
	Female	9	116	125	2	42	44
Cough Data	No. of voluntary coughs	187	1863	2050	86	612	698
	No. of spontaneous coughs	175	1037	1212	37	358	395
	Total no. of coughs	362	2900	3262	123	970	1093
	Average no. of coughs/patient	8.42	9.03	8.96	9.46	9.51	9.50

B. Results using Manual Segmentation on Dataset A

1) Results using Raw Features

Croup and non-croup classification results using raw features, or without feature selection, on manually segmented cough files are discussed in this subsection. This includes the results for log MFCC, linear MFCC, CIF, log MFCC + CIF, and linear MFCC + CIF. The classification results for these features with LRM and SVM classification methods are given in Table II-A and Table II-B, respectively.

The number of non-croup patients is about 7.5 times more than the number of croups patients which makes the accuracy highly inclined towards the specificity. However, the average classification performance (average of the sensitivity and specificity values) of the SVM classifier is generally superior to the LRM classifier and this is especially true when features were combined.

Looking at the classification performance of the individual features, when compared to log MFCC, the sensitivity value obtained using linear MFCC is significantly better with both LRM and SVM classification methods. With LRM classification, the sensitivity increases from 83.72% to 93.02% (+9.30%) while the specificity drops slightly from 82.55% to 82.24% (–0.31%). Similarly, with SVM classification, the sensitivity increases from 81.40% to 95.35% (+13.95%) while there is a marginal decline in specificity from 91.59% to 88.79% (–2.80%). The average classification performance of linear MFCC is significantly superior to log MFCC with both classifiers.

The CIF is seen to be more useful than log MFCC in differentiating between croup and non-croup cough sound with both classification methods. With LRM classification, the increase in sensitivity and specificity from log MFCC to CIF are +9.30% and +1.56%, respectively. Similarly, the increase in sensitivity and specificity with SVM classification are +4.65% and –0.31%. As such, the average performance of the proposed time-frequency image feature is significantly better

than the conventional MFCC.

Comparing with linear MFCC, while the average classification performance of CIF is slightly better with LRM classification, the average classification performance of linear MFCC is significantly better than CIF with SVM classification. At a sensitivity of 95.35% and specificity of 89.72%, this is also the best overall classification performance of all the individual features considered in this work. Also, in general, the average classification performance of the individual features is seen to be better with SVM than LRM.

A similar conclusion can also be drawn with feature vector combination of linear MFCC + CIF over log MFCC + CIF. SVM is once again determined to be superior for raw feature classification when compared to LRM. With SVM classification, from log MFCC + CIF to linear MFCC + CIF, the sensitivity value increases from 88.37% to 97.67% (+9.30%) and the specificity value increases from 91.59% to 92.21% (+0.62%).

2) Results Using Feature Selection

Results using feature selection are presented in this subsection. Due to the significantly better overall performance of linear MFCC over log MFCC in distinguishing between croup and non-croup patients using raw features, only linear MFCC is considered here. Similarly, the feature combination of linear MFCC + CIF is considered here over log MFCC + CIF due to its significantly better overall performance with raw features.

In Table III-A, we present the classification results using *p*-value feature selection for the LRM classifier. The improvement in sensitivity and specificity values (and the average improvement) for linear MFCC, CIF, and linear MFCC + CIF classified after *p*-value feature selection against raw features are as follows: +2.33% and +4.99% (+3.66%), 0% and +0.63% (+0.32%), and 0% and +5.92% (+2.96%). As such, there is some degree of improvement in the classification performance for all feature sets when compared

TABLE II-A
PATIENT CLASSIFICATION RESULTS USING LRM ON MANUALLY SEGMENTED COUGHS

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Log MFCC	83.72	82.55	82.69	39.13	97.43	0.44
Linear MFCC	93.02	82.24	83.52	41.24	98.88	0.49
CIF	93.02	84.11	85.16	43.96	98.90	0.52
Log MFCC + CIF	88.37	75.70	77.20	32.76	97.98	0.37
Linear MFCC + CIF	97.67	79.75	81.87	39.25	99.61	0.47

TABLE II-B
PATIENT CLASSIFICATION RESULTS USING SVM ON MANUALLY SEGMENTED COUGHS

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Log MFCC	81.40	91.59	90.38	56.45	97.35	0.61
Linear MFCC	95.35	89.72	90.38	55.41	99.31	0.65
CIF	86.05	91.28	90.66	56.92	97.99	0.63
Log MFCC + CIF	88.37	91.59	91.21	58.46	98.33	0.65
Linear MFCC + CIF	97.67	92.21	92.86	62.69	99.66	0.72

to results with raw features.

Classification results using backward sequential feature selection with the LRM classifier for the same features as in Table III-A are given in Table III-B. When compared to the results using raw features, the improvement in sensitivity and specificity values (and the average improvement) for linear MFCC, CIF, and linear MFCC + CIF are as follows: +4.65% and +4.68% (+4.67%), +2.33% and +4.05% (+3.19%), and 0% and +9.04% (+4.52%). The difference in average improvement using the backward sequential feature selection method over the p-value feature selection approach for linear MFCC, CIF, and linear MFCC + CIF are +1.01%, +2.87%, and +1.56%, respectively. As such, backward sequential feature selection is seen to be more effective than p-value feature selection method for the LRM classifier.

In Table III-C, results for linear MFCC, CIF, and linear MFCC + CIF using backward sequential feature selection applied to the SVM classifier are presented. The improvement in sensitivity and specificity values for linear MFCC, CIF, and linear MFCC + CIF over the SVM results using raw features are as follows: +2.32% and +1.25% (+1.79%), +4.65% and +1.24% (+2.95), and 0% and +4.36% (+2.18%). As such, backward sequential feature selection is also seen to improve the classification performance of the SVM classifier over the results using raw features. While the average improvement for each feature or feature set is lower than the LRM classifier, the SVM classifier is generally is seen to give a better overall performance. With sensitivity and specificity value of 97.67% and 96.57%, respectively, the best classification performance is once again achieved using the feature combination of linear MFCC + CIF using SVM classification.

C. Results using Auto Segmentation on Dataset A

The auto segmentation algorithm described in Appendix A was trained, validated, and tested on Dataset A. Dataset A was divided in training/validation and test sets, 153 patients and 211 patients, respectively, for this purpose.

The auto segmentation validation and testing results are given in Table IV. The start and end points of the auto segmented coughs were compared with manually segmented coughs and the auto segmentation accuracy for each patient was computed. Details on this comparison procedure can be found in our earlier work in [24]. The values given in Table IV are the mean and standard deviation for all patients and croup and non-croup patients. We do not report the specificity values here since non-cough events are not of interest and, therefore, were not manually segmented for this computation.

On the validation set, on average, 89.79% of coughs per patients are correctly segmented which reduces to 85.54% of coughs per patients on the test set. For croup patients, an average of 83.57% of coughs per patient is correctly segmented on the validation set and, interestingly, increasing to 85.48% of coughs per patient on the test set. For non-croup patients, an average of 90.42% of coughs per patient are correctly segmented on the validation set decreasing to 85.55% on the test set. In general, a good correlation is observed between the results on the validation and test sets. Also, a mean accuracy of $>85\%$ on the test set indicates the robustness of the auto segmentation algorithm.

Next, croup and non-croup patient classification accuracy values using auto segmented coughs are presented. Results for linear MFCC, CIF, and linear MFCC + CIF using LRM and SVM classification methods are given in Table V-A and Table V-B, respectively. For linear MFCC, we used the same number of mel-filters as in manual segmentation. Similarly, we used same number of blocks for the CIF. However, this time we trained the models using auto segmented files and then applied backward sequential feature selection.

Generally, there is a slight decline in the sensitivity and specificity values over those achieved using manually segmented cough files. Looking at the results using the best cough feature set, linear MFCC + CIF, the difference in sensitivity and specificity values against the results using

TABLE III-A
PATIENT CLASSIFICATION RESULTS USING LRM ON MANUALLY SEGMENTED COUGHS (P-VALUE FEATURE SELECTION)

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Linear MFCC	95.35	87.23	88.19	50.00	99.29	0.59
CIF	93.02	84.74	85.71	44.94	98.91	0.53
Linear MFCC + CIF	97.67	85.67	87.09	47.73	99.64	0.57

TABLE III-B
PATIENT CLASSIFICATION RESULTS USING LRM ON MANUALLY SEGMENTED COUGHS (BACKWARD SEQUENTIAL FEATURE SELECTION)

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Linear MFCC	97.67	86.92	88.19	50.00	99.64	0.60
CIF	95.35	88.16	89.01	51.90	99.30	0.61
Linear MFCC + CIF	97.67	88.79	89.84	53.85	99.65	0.64

TABLE III-C
PATIENT CLASSIFICATION RESULTS USING SVM ON MANUALLY SEGMENTED COUGHS (BACKWARD SEQUENTIAL FEATURE SELECTION)

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Linear MFCC	97.67	90.97	91.76	59.15	99.66	0.69
CIF	90.70	92.52	92.31	61.90	98.67	0.69
Linear MFCC + CIF	97.67	96.57	96.70	79.25	99.68	0.86

manual segmentation are –4.81% and –5.04% with LRM classification and –2.43% and –6.57% with SVM classification, respectively. With a sensitivity of 95.24% and specificity of 90.00%, the best overall results are once again achieved using SVM classification.

The decline in the sensitivity and specificity values are due to the differences in the two models which in turn is due to the disparities introduced by the auto segmentation algorithm. When compared to the manually segmented coughs, three key differences with the automatically segmented coughs are loss of cough events as non-cough events (as per the cough segmentation sensitivity values in Table IV), detection of non-cough events as cough events (as per the PPV in Table IV), and the differences in the start and end points of the coughs.

D. Prospective Testing Based on Auto Segmentation of Dataset B

Patient classification results on Dataset B (the prospective dataset) are discussed in this subsection. We only considered auto segmented cough sound signals for this purpose since this is how a fully automated croup diagnosis system should operate in practice.

The model that was subjected to the prospective testing was trained on the training and validation dataset (Dataset A) and all model parameters were fixed at the end of the training procedure. We call this model the *Final Croup Model*. The *Final Croup Model* training followed the same procedure applied in Section V-C except that the whole of the Dataset A was used here. This gave us one single *Final Croup Model* to test on the prospective data set (Dataset B).

Croup and non-croup patient classification results on Dataset B are given in Table VI for both LRM and SVM classification methods. Only the best results for LRM and SVM classification methods are presented here which is using the feature set of linear MFCC + CIF.

The sensitivity and specificity values are 92.31% and 78.43% using LRM classification and 92.31% and 85.29% using SVM classification, respectively. The sensitivity value using the two classifiers is same. The specificity value, however, is about 7% higher with SVM classification. Performance on the prospective set is close to that on the training/validation set (within 0.55-5.32%) indicating the generalization capability of the *Final Croup Model*.

For the SVM results, the 95% confidence interval is 77.84% to 100% for the sensitivity and 78.42% to 92.16% for the specificity. In addition, the ROC for the SVM classifier is shown in Fig. 5. As with the validation set, the PPV values are low which can be attributed to the significantly less number of croup subjects compared to non-croup subjects.

VI. DISCUSSION AND CONCLUSION

Based on the performance of our technology on both the training/validation and prospective data sets, we conclude that it is indeed possible to accurately and objectively diagnose croup using cough sounds alone.

It is possible to augment cough-based features with simple symptoms observable by parents targeting further improvement in performance. Results of our early feasibility studies along these lines on a larger dataset is available in [22].

The results we obtained in this paper corroborate our previous finding that the combination of multiple features improves the diagnostic performance of our models. The feature combination of linear MFCC and CIF produced the best classification performance on our real-world cough sound

TABLE IV
VALIDATION AND TESTING RESULTS FOR THE AUTO SEGMENTATION ALGORITHM ON DATASET A

| | Validation Results | | Testing Results | |
	Sensitivity	PPV	Sensitivity	PPV
All Patients	89.79	80.55	85.54	82.68
Croup Patients	83.57	77.69	85.48	77.18
Non-croup Patients	90.42	80.84	85.55	83.50

TABLE V-A
PATIENT CLASSIFICATION RESULTS USING LRM ON AUTO SEGMENTED COUGHS (BACKWARD SEQUENTIAL FEATURE SELECTION)

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Linear MFCC	90.48	84.06	84.81	42.70	98.53	0.50
CIF	83.33	90.00	89.23	52.24	97.63	0.58
Linear MFCC + CIF	92.86	83.75	84.81	42.86	98.89	0.51

TABLE V-B
PATIENT CLASSIFICATION RESULTS USING SVM ON AUTO SEGMENTED COUGHS (BACKWARD SEQUENTIAL FEATURE SELECTION)

	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
Linear MFCC	92.86	85.94	86.74	46.43	98.92	0.55
CIF	88.10	91.56	91.16	57.81	98.32	0.65
Linear MFCC + CIF	95.24	90.00	90.61	55.56	99.31	0.65

TABLE VI
PATIENT CLASSIFICATION RESULTS ON DATASET B FOR LINEAR MFCC + CIF MODEL DEVELOPED USING AUTO SEGMENTED COUGHS AND BACKWARD SEQUENTIAL FEATURE SELECTION ON DATASET A

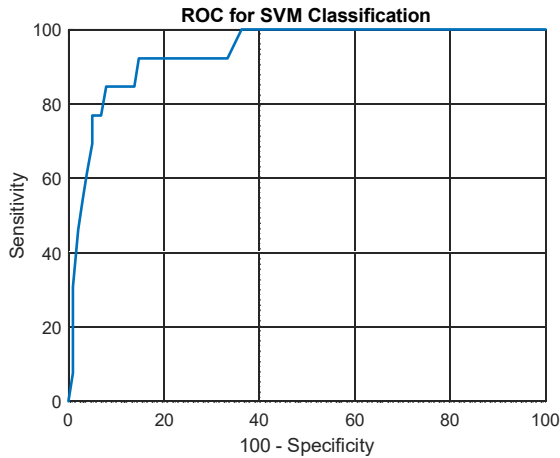|	Sensitivity	Specificity	Accuracy	PPV	NPV	κ
LRM	92.31	78.43	80.00	35.29	98.77	0.41
SVM	92.31	85.29	86.09	44.44	98.86	0.53

Fig. 5. ROC using SVM classification on the test dataset.

signal database with clinically diagnosed respiratory diseases. Our results also lead to the conclusion that feature dimension reduction techniques are useful in improving the model performance.

The work we reported in this paper was based on two different classifiers (a linear LRM and a non-linear SVM). The SVM was determined to be more accurate than LRM in classifying the cough sound signals, particularly with the bio-mimicking features we explored here. More sophisticated classifier models and features can be explored in the future with larger data sets.

APPENDIX A

This work builds on our earlier work on automatic cough segmentation using Time Delay Neural Network (TDNN) in [24]. In this work, we created a TDNN by implementing autoencoders [25] in the hidden layers to design a Time Delay Deep Neural Network (TD-DNN) [25, 26]. An autoencoder is a feed forward neural network trained to reproduce its input at the output [25]. The hidden layers in an autoencoder symbolize a code which can be used to represent the input data. We divided the whole process of training a TD-DNN into three stages.

Stage 1: Input feature vector – The audio data stream was divided into contiguous sub-blocks of duration 20 ms. The following mathematical features were computed from each sub-block: 34 MFCCs, the "pitch-ness" coefficient (defined as the ratio of second peak to the first peak of the autocorrelation of the data in sub-block), Shannon's entropy, and zero-crossing rate. Of these 37 features, 25 features were determined as having a strong association with cough sounds using analysis of variance hypothesis test. Thus, the feature
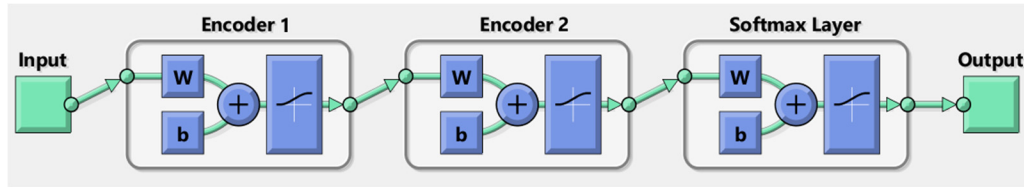
Fig. 6. The TD-DNN architecture trained to automatically segment cough sound signals. Size of the network is as follows: input feature vector = 125 dimensional, autoencoder 1 = 10 neurons, autoencoder 2 = 5 neurons, and softmax output layer = 1 neuron.

vector for the uth subject in sub-block z, F_z^u, consists of 25 selected features.

Stage 2: TD-DNN architecture and layer-by-layer network training – The TD-DNN architecture has an input layer (L_i), two hidden layers (L_{h1} and L_{h2}), and an output layer (L_o) as shown in Fig. 6. The hidden and output layers (L_{h1}, L_{h2}, and L_o) are independently trained using the output from the previous layer.

The number of neurons in L_i depends on the size of the input feature vector and number of time delays. For this work, we used a time delay of 5, that is, to classify z^{th} sub-block as cough or non-cough, feature vectors $\{F_{z-2}^u \ F_{z-1}^u \ F_z^u \ F_{z+1}^u \ F_{z+2}^u\}$ are used as inputs to the TD-DNN. Therefore, the input layer L_i is 125 dimensional.

The first hidden layer L_{h1} is an autoencoder of size 10 and is trained using the input layer feature vector. The second hidden layer L_{h2} is the second autoencoder of size 5 and is trained using the encoded output of the first autoencoder. The output layer L_o is a softmax function with 1 neuron and is trained using the encoded output of the second autoencoder.

Stage 3: Fine-tuning stage – In this stage, all trained layers from stage 2 were connected together to create a stacked TD-DNN. The TD-DNN was then retrained with limited number of training epochs (maximum 200) to fine-tune the TD-DNN network parameters.

A detailed description of the method is out-of-scope of this paper and will be reported elsewhere.

REFERENCES

[1] WHO, *Management of the child with a serious infection or severe malnutrition: guidelines for care at the first-referral level in developing countries.* World Health Organization, 2000.
[2] A. O. Segal, E. J. Crighton, R. Moineddin, M. Mamdani, and R. E. Upshur, "Croup hospitalizations in Ontario: A 14-year time-series analysis," *Pediatrics,* vol. 116, no. 1, pp. 51-5, 2005.
[3] J. Charles, H. Britt, and S. Fahridin, "Croup," *Australian Family Physician,* vol. 39, no. 5, pp. 269-269, 2010.
[4] C. L. Bjornson and D. W. Johnson, "Croup," *The Lancet,* vol. 371, no. 9609, pp. 329-339, 2008.
[5] C. L. Bjornson and D. W. Johnson, "Croup in children," *CMAJ: Canadian Medical Association Journal,* vol. 185, no. 15, pp. 1317-1323, 2013.
[6] "Diagnosis and management of croup," Toward Optimized Practice, 2008.
[7] R. V. Sharan and T. J. Moir, "Cochleagram image feature for improved robustness in sound recognition," in *Proceedings of the IEEE International Conference on Digital Signal Processing (DSP)*, Singapore, 2015, pp. 441-444.
[8] V. Tyagi and C. Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 2005, pp. 529-532.
[9] R. Sarikaya and J. H. Hansen, "Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition," in *EUROSPEECH-2001*, Aalborg, Denmark, 2001, pp. 687-690.
[10] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Annals of Biomedical Engineering,* vol. 41, no. 11, pp. 2448-2462, Nov 2013.
[11] K. Kosasih, U. R. Abeyratne, V. Swarnkar, and R. Triasih, "Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis," *IEEE Transactions on Biomedical Engineering,* vol. 62, no. 4, pp. 1185-1194, 2015.
[12] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PLOS ONE,* vol. 11, no. 9, pp. 1-20, 2016.
[13] R. V. Sharan, U. R. Abeyratne, V. R. Swarnkar, and P. Porter, "Cough sound analysis for diagnosing croup in pediatric patients using biologically inspired features," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jeju Island, South Korea, 2017, pp. 4578-4581.
[14] D. O'Shaughnessy, *Speech communication: human and machine.* Addison-Wesley Pub. Co., 1987.
[15] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*. vol. 83, Y. Cazals, L. Demany, and K. Horner, Eds. Pergamon, Oxford, 1992, pp. 429-446.
[16] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *Journal of the Acoustical Society of America* vol. 87, no. 6, pp. 2592-2605, Jun 1990.
[17] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer, Technical Report 35, 1993.
[18] M. Slaney, "Auditory Toolbox for Matlab," Interval Research Corproation, Technical Report 1998-010, 1998.
[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.
[20] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, Pittsburgh, Pennsylvania, USA, 1992, pp. 144-152.
[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, *et al.*, *The HTK book (for HTK version 3.4).* Cambridge University: Engineering Department, 2009.
[22] ResApp Health Limited. (2017). *ResApp provides updated australian paediatric study results*. [Accessed: 15 Jan, 2017] Available: https://www.resapphealth.com.au/
[23] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012, pp. 2951-2959.
[24] Y. A. Amrulloh, U. R. Abeyratne, V. Swarnkar, R. Triasih, and A. Setyati, "Automatic cough segmentation from non-contact sound recordings in pediatric wards," *Biomedical Signal Processing and Control,* vol. 21, pp. 126-136, 2015.
[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, no. 5786, pp. 504-507, 2006.
[26] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks,* vol. 61, pp. 85-117, 2015.