Hindawi Publishing Corporation
Modelling and Simulation in Engineering
Volume 2012, Article ID 983147, 10 pages
doi:10.1155/2012/983147

Research Article

Cry-Based Classification of Healthy and Sick Infants Using Adapted Boosting Mixture Learning Method for Gaussian Mixture Models

Hesam Farsaie Alaie and Chakib Tadj

École de Technologie Supérieure, Université du Québec, 1100 rue Notre-Dame Ouest, Montréal, QC, Canada H3C 1K3

Correspondence should be addressed to Hesam Farsaie Alaie, hesam.farsaei@gmail.com

Received 28 August 2012; Revised 21 November 2012; Accepted 30 November 2012

Academic Editor: Shinsuke Hara

We make use of information inside infant's cry signal in order to identify the infant's psychological condition. Gaussian mixture models (GMMs) are applied to distinguish between healthy full-term and premature infants, and those with specific medical problems available in our cry database. Cry pattern for each pathological condition is created by using adapted boosting mixture learning (BML) method to estimate mixture model parameters. In the first experiment, test results demonstrate that the introduced adapted BML method for learning of GMMs has a better performance than conventional EM-based reestimation algorithm as a reference system in multipathological classification task. This newborn cry-based diagnostic system (NCDS) extracted Mel-frequency cepstral coefficients (MFCCs) as a feature vector for cry patterns of newborn infants. In binary classification experiment, the system discriminated a test infant's cry signal into one of two groups, namely, healthy and pathological based on MFCCs. The binary classifier achieved a true positive rate of 80.77% and a true negative rate of 86.96% which show the ability of the system to correctly identify healthy and diseased infants, respectively.

1. Introduction

Crying is the first sound the baby makes when he enters the world outside of his mother's womb, which is a very positive sign of a new healthy life. Infants cry for the same reason that adults talk, that is, to let others know about their needs or problems. Since crying is all a baby can do to express any discomfort, it seems that this multimodal signal carries a lot of information about him. In early studies of the infant cry analysis, the acoustic structure of infant crying was analyzed, and some of the important variables controlling the production of their cries were described [1]. After the cry analysis on infants with various diseases, in some cases it has been noticed that there are fixed cry attributes, which are rarely seen in cries of healthy infants. Instead, these attributes occur frequently in cries of infants with diseases [1–3]. Therefore the concealed information contained within a cry signal could clarify the infant's present psychological condition. Acoustic analysis of the infants' cry signal helps to measure these parameters quantitatively to perform a comparison between healthy and ill states. Since infants' cry could be changed from normal to abnormal by diseases or deformities which have the ability to produce an ill effect on the central nervous system, the oral cavities, or respiratory organs, our goal is to develop an NCDS to classify infants with different kind of physiological conditions.

Generally, sounds can be represented in multiple ways. In feature representation method, the features selection step relies heavily on good understanding of the problem. There are some literatures on defining and using different cry characteristics and frequency features which distinguish between a healthy infant's cry and that of infants with asphyxia, brain damage, hyperbilirubinemia, Down's syndrome, and mothers who were drug abused during pregnancy [1, 4]. Human speech features characteristics such as linear prediction coefficients (LPCs), MFCCs, and fundamental frequency and formants are studied in previous works [2, 5–9]. The work presented by Hariharan et al.

employed wavelet packet transform (WPT) to compute subband energy and entropy features from wavelet packet coefficients [10]. The goal is given a set of exemplary patterns for C different pathological infant's cry classes, to construct a function such that when presented with a new feature from an infant's cry belonging to class i, the function will recognize the correct index pathology class. In recent years, several machine learning algorithms such as artificial neural network (ANN), radial basis function (RBF), and probabilistic neural network (PNN) have demonstrated their ability to recognize cry patterns and make intelligent decision based on available training database [8, 10, 11]. Furthermore, hybrid systems in which classification methods are combined under several approaches like bagging [12], boosting [12], majority voting, and staking were examined in order to achieve better final results than the case where a single classifier is run [5–7].

In this paper we make use of extracted MFCCs from an infant's cry signal to diagnose pathological conditions and specific diseases which have not been previously studied such as "Coarctation of Aorta" and "Tetralogy of Fallot" by drawing support from collected cry database. As we mentioned earlier, there exist a large number of approaches to do the modeling and the classification tasks. We will focus on GMMs, which are the most successful classifiers in use for audio data when their temporal structure is not important [13]. This paper employs GMMs to introduce a classification technique in the field of statistical learning theory which uses adapted BML method to train mixture models for modeling of infants' cry signals. The BML method [14] presents three key advantages: (1) Add new components into the direction that largely increases the objective function, (2) Decrease sensitivity to initial parameters, and (3) Estimate the optimum number of components, unlike the conventional EM-based reestimation algorithm. Partial and global updating methods were used in model parameters estimation processes in order to speed the learning process up and converge to more robust and reliable estimation of a new mixture component. Another advantage of the adapted BML method was that it used Bayesian inference criterion (BIC) [15] for model selection. It is partially based on the likelihood function, but to avoid overfitting there is also a penalty term.

This paper is organized as follows: in Section 2 we give a brief review of GMM, its role in cry-pattern classifier, and advantages of adapted BML method. Section 3 explains the different parts of the NCDS and how it identified infants. In Section 4, results of experiments in both multipathological and binary classification tasks are reported, and in Section 5 a follow-up analysis of the results and conclusion are presented to finalize the paper.

2. GMMs for Cry-Pattern Classification

GMMs are a special case of hidden Markov models (HMMs) which pay more attention to the temporal structure of a sound and have proven to be invaluable tools in areas such as speech recognition [13]. Compared to the human speech which is modeled by hidden Markov models with finite states, cry signal has just a single state to be considered. Moreover, GMM has the ability to form smooth approximations from arbitrarily shaped densities, and it has proved to be an effective probabilistic model for applications such as biometric systems, most notably in speaker recognition systems and speaker identification [16] due to its capability of representing a large class of sample distributions. In a cry-based diagnosis system there is no chance to train the classifier with a specific individual compared to what happens in speaker-dependent automatic speech recognition (ASR) systems. Therefore we should use our available cry database to create a more general model for each available pathological condition in order to fine-tune the pathology detection of test infants. Like ASR systems in the learning phase, there are two main parts [17]: first, some cry features are selected, and then patterns are created from these features. Second, the cry-pattern classifier works according to the just created cry patterns to recognize physiological conditions on the newborn infants. The proposed cry-modeling method trained the GMM classifier from feature streams. It means that cry signals coming from either healthy or pathological classes were modeled by a separate pool of Gaussians using extracted feature vectors. Adapted BML was used to learn mixture models in an incremental and recursive manner in which, unlike EM-based re-estimation algorithm, the final model was less sensitive to initial parameters. It might, therefore, converge to a better optimal point. According to the boosting theory, upper bound on training error rate is analytically minimized and the margin of all training samples is increased. Moreover, in the preceding paper [14] the strong point of BML method shows that a new Gaussian component is estimated in each step according to the functional gradient of the objective function. Therefore each new component is always added to the direction that increases the objective function the most. In this paper adapted BML has been described for the log-likelihood function of the mixture model over training data as our objective function.

3. Newborn Cry-Based Diagnosis (NCDS) System

3.1. Cry Database. The number of labeled data used during training phase has a leading role in performance of the classifier. For example, in case of having small number of cry samples for the pathologies of interest, the resulting trained classifier may be too specific to be generalized to unseen infant's cry signal. Cry database collecting is still in progress and up to now, the following two kinds of newborn infants are considered in the database:

(1) healthy newborn infants, both full term and premature,

(2) sick newborn infants, both full term and premature, with specific selected pathologies.

The imbalanced learning problem [18] is concerned with the performance of learning algorithms. In the presence of underrepresented data and skewed class distribution this problem arises, which is a direct result of the nature of the data space present in the cry database. Table 1 shows the list

Table 1: Cry database.

Infants	State	Pathologies	Number
Full term	Healthy	N/A	38
	Pathologies	Bovine protein allergy	13
		Tetralogy of Fallot	5
		Thrombosis in the vena cava	13
Premature	Healthy	N/A	25
	Pathologies	Tetralogy of Fallot	9
		Cardio complex	14
		X chromosomal abnormalities	9
		Coarctation of aorta	10

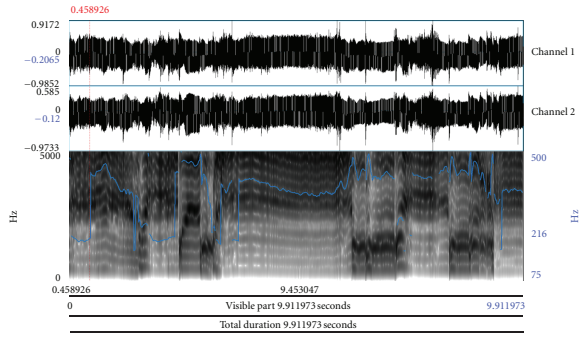

Figure 1: Time domain representation of a cry signal.

of different pathologies and the number of available samples in each class.

3.2. Preprocessing and Feature Extraction. In data collection step we have used 2-channel recorder with 44.1 kHz sampling rate. The time domain representation of one cry signal in two channels is shown in Figure 1. In preprocessing step it converted into one-channel signal using mean value function.

In the acoustical analysis step, the resulting wave was cleared of any silence region or external unwanted sounds as for the nurse or pediatrician voices, then normalized, and split into frames. Next, MFCCs were extracted. The vocal tract shape generally changes slowly with time and tends to be constant over short intervals. A reasonable approximation to guarantee reproducibility is to analyze the signal into a sequence of millisecond-time frames, where each frame is represented by a single feature vector describing the average spectrum for a short time interval [19]. In reading about application of frequency feature analysis of speech signals, it is common practice to preemphasize the signal prior to compute the parameters by applying the first order difference equation $s'(n) = s(n) - \alpha s(n-1)$ to the sample sequence $\{s(n), n = 1, \ldots, N\}$ in each window of length N. The z-transform of the filter is $P(z) = 1 - \alpha z^{-1}$. Deller [20] had earlier referred to the reasons behind employing a preemphasis filter which are twofold: first, it is due to canceling spectral effects of one of the glottal poles, and the second reason is to prevent numerical instability. Gray Jr. and Markel [21] and Makhoul and Viswanathan [22] have worked with an optimal value of α given by $\alpha = r_s(1;m)/r_s(0;m)$ in the sense of MSE, where $r_s(\eta;m)$ is the usual short-term autocorrelation sequence for the frame ending at m with the parameter η corresponding to the autocorrelation lag. One estimator is given by [20]:

$$r_s(\eta;m) = \frac{1}{N}\sum_{n=-\infty}^{\infty} s(n)w(m-n) \times s(n-|\eta|)w(m-n+|\eta|), \quad (1)$$

where $w(n)$ is a window of length N. For voiced frames the optimal value is near unity, whereas for unvoiced frames it is small ($\alpha \approx 0$). Therefore it should not be performed on unvoiced speech and in voiced frames; it is taken in the range $0.9 \le \alpha \le 1$ which introduces a zero near $\omega = 0$, and a $6 - db$ per octave shift on the spectrum. We use a common value for this factor which is 0.97 [23–26]. In the next step, L MFCCs per frame were extracted just for those voiced frames. The sequence of feature vectors describing the average spectrum for a short time interval can be presented as a matrix which acts like a pattern in the classification phase.

In all related practical applications, the short terms or frames should be utilized which implies that the signal characteristics are uniform in the region. Therefore the selected portion of the signal has to be short enough to be stationary. Temporal properties can be assumed fixed over time intervals on the order of 10–30 msec [27]. Prior to any frequency analysis, the Hamming windowing is necessary to reduce any discontinuities at the edges of the selected region. Generally a longer window will tend to produce a better spectral picture of the signal while the window is completely within a stationary region, whereas a shorter window will tend to resolve events in the signal better in time. This tradeoff is sometimes called the spectral temporal resolution tradeoff [20]. A common choice for the value of the window length (10–30 msec) [17, 27–29] is normally larger than the frame rate. For example, the typical value for the window length in HTK [23] is 25 msec. MFCCs are obtained by applying the discrete cosine transform (DCT) to the output of the mel-filters. The difference from the real cepstrum is that a nonlinear frequency scale is used, which approximates the behavior of the auditory system [27]. A filter bank with K filter is defined, where all these

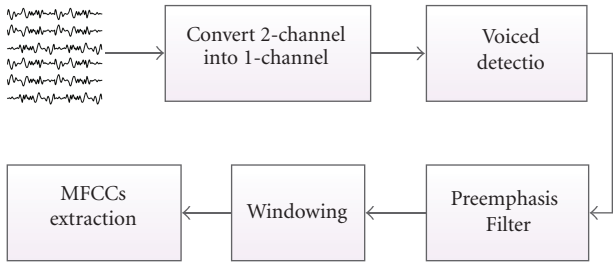

Figure 2: Preprocessing steps.

triangular filters compute the average spectrum around each center frequency with increasing bandwidth. An acoustic representation using MFCCs is often referred to as a "mel cepstrum." After performing fast Fourier Transform (FFT) on each windowed frame, MFCCs are calculated using the following DCT [13]:

$$c_n = \sqrt{\frac{2}{K}} \sum_{i=1}^{K} \log S_i \times \cos\left(n\left(i - \frac{1}{2}\right)\frac{\pi}{K}\right), \quad n = 1, 2, \ldots L, \tag{2}$$

where K is the number of subbands, L is the desired length of cepstrum, and S_i, $i = 1, 2, \ldots K$ $(1 \le i \le K)$ represents the filter bank energy after passing through the triangular band pass filters. We use a set of 20 triangular windows $(K = 20)$ which is utilized in a common approach to simulate critical-band filtering [30, 31], whose energy outputs are designed S_i, $i = 1, 2, \ldots, 20$. We will discuss the choice of the parameter L later. Figure 2 shows all the pre-processing steps from the cry-recording step until the extraction of the MFCCs.

3.3. Adapted BML Method for GMMs. The main idea of the boosting method in machine learning is that instead of always treating all data points as equal, component classifiers should specialize in certain samples. In particular, if a sample is hard to classify and problematic for the existing classifier, more components should focus on it. Compared to other learning mixture models [32–34], BML method [14] has a great privilege to add new mixture components in such a way that has the greatest improvement in the predefined objective function, $\mathcal{C}(F_k)$. Moreover, this method is less sensitive to initial parameters, resulting in better optimal point in the convergence process. The whole cry-pattern classification approach comprised a classification scheme using GMMs that classify patterns created by extracted real-valued frequency domain features. A GMM, $F_K(X)$ with K Gaussian components, and given feature vector X can be represented as

$$F_K(X) = \sum_{k=1}^{K} \pi_k f_k(X), \tag{3}$$

with the restriction that $0 \le \pi_k \le 1$ for $k = 1, \ldots, K$ and

$$\sum_{k=1}^{K} \pi_k = 1, \tag{4}$$

where π_k and $f_k(X)$ are mixture proportions and distributions of the kth component, respectively. The kth multivariate Gaussian component of a D-dimensional feature vector X can be written in the following notation:

$$f_k(X \mid \Phi_k) = \mathcal{N}(X; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \times \exp\left(-\frac{1}{2}(X - \mu_k)^{Tr}{\Sigma_k}^{-1}(X - \mu_k)\right), \tag{5}$$

where $\Phi_k = \{\mu_k, \Sigma_k\}$ are the mean and covariance parameters for the kth component, and A^{Tr} represents the transpose of matrix A. The model commences with one mixture and learns gradually by adding a new mixture component on each step. According to the defined objective function $\mathcal{C}(\cdot)$, each adding component process should satisfy the inequality $\mathcal{C}(F_k) > \mathcal{C}(F_{k-1})$. All the reestimation formulas to update Gaussian distribution parameters can be computed afterwards. Log-likelihood function of the mixture model over all training feature data has a vital role in aforementioned formulas. For example, the predefined objective function of the mixture model over all training feature can be computed using the equation below:

$$\mathcal{C}\left(F_j\right) = \sum_{t=1}^{T_j} \log F_j(X_t), \tag{6}$$

where T_j is the number of training feature vectors X in jth pathology class. Iterative re-estimation formulas for model parameters $\Phi_k^{(n+1)} = \{\mu_k^{n+1}, \Sigma_k^{n+1}\}$ at the $(n+1)$th iteration can be evaluated as follows:

$$\begin{aligned} w^n(X_t) &= \frac{f_k\left(X_t \mid \Phi_k^{(n)}\right)}{F_{k-1}(X_t \mid \Psi_{k-1})}, \\ {\mu_k}^{n+1} &= \frac{\sum_{t=1}^{T_j} w^n(X_t) \cdot X_t}{\sum_{t=1}^{T} w^n(X_t)} \\ &= \sum_{t=1}^{T_j} \gamma_t\left(\Phi_k^{(n)}\right) \cdot X_t, \\ \Sigma_k^{n+1} &= \frac{1}{\sum_{t=1}^{T_j} w^n(X_t)} \\ &\quad \times \sum_{t=1}^{T_j} w^n(X_t)\left(X_t - \mu_k^{n+1}\right)\left(X_t - \mu_k^{n+1}\right)^{Tr}, \end{aligned} \tag{7}$$

where $\Phi_k^{(n)} = \{\mu_k^n, \Sigma_k^n\}$ denotes the mean vector and covariance matrix for kth Gaussian component at the nth iteration, and $\Psi_k = \{\Phi_k, \Psi_{k-1}\}$. Note how $w^n(X_t)$ represents a weight assigned to feature vector X_t after nth iteration. As you can see according to this weighting function it is clear to understand that feature vectors with lower probability by current model are given larger weights than those with more probability. Therefore, the new Gaussian component f_k focuses on those features which are poorly modeled by the current model F_{k-1}, in much the same way as other boosting algorithms [35, 36].

3.4. Initialization of Sample Weights. There is a problem with the initialization values of weights based on boosting theory which can be computed by using equation below:

$$w^0(X_t) = \frac{1}{F_{k-1}(X_t \mid \Psi_{k-1})}, \tag{8}$$

where $\Psi_{k-1} = \{\Phi_{k-1}, \Psi_{k-2}\}$. The dynamic range of F_{k-1} is large in a way that it could be dominated by only a few number of samples with low probability or outliers. We use the so-called "Weight decay" method [37] to overcompensate for the low probability by smoothing sample weights based on power scaling:

$$w^0(X_t) = \left(\frac{1}{F_{k-1}(X_t \mid \Psi_{k-1})}\right)^p, \quad 0 < p < 1, \tag{9}$$

where p is a decay parameter or an exponential scaling factor. In the second method the idea of sampling boosting in [35] is applied to form a subset of training feature vectors according to the mean and variance values of the decayed weights. Afterwards, vectors contained in the just created subset are utilized with equal weights to estimate the new component parameters. Assume $\overline{M}$ and σ^2 denote the mean and variance of weights calculated in (8) as defined below:

$$\begin{aligned} \overline{M} &= \text{mean}\{\log w^0(X_t)\}, \\ \sigma^2 &= \text{variance}\{\log w^0(X_t)\}. \end{aligned} \tag{10}$$

Then, the aforementioned subset with large weights is selected as described below:

$$X_{\text{sub}} = \left\{X_t \mid \log w^0(X_t) > \overline{M} + \beta\sigma\right\}, \tag{11}$$

where β is a linear scaling factor to control the size of subset X_{sub}.

3.5. Process of Adding a New Component. In the adding process the part of training vectors in which $f_k(X)$ had a higher value than the reminder of the mixture model, denoted by $F_k - \{f_k\}$, was selected. Then this subset of data X_{sub} should be modeled by a small GMM consisting of two Gaussian components called f_k^* and f_{K+1}. The initial component came from the EM-based re-estimation algorithm, and then the second component and its weight were estimated based on BML method and line search, respectively. We considered the estimated component—the second one—as an initial component and ran BML method again. This process continues repeatedly, until it reached the optimal maximum log-likelihood estimate of parameters over X_{sub}. This procedure for finding the best two new components f_{K+1} and f_k^* continued for $k = 1, \ldots, K$. Amongst all the created K mixture models, denoted by F_{K+1}, the one that gave the highest value of the objective function (same as log-likelihood value) was selected and added to the mixture by adjusting its weight.

3.6. Partial and Global Updating. In the previous step, instead of finding the new mixture weight from the line search below:

$$\pi_k^* = \underset{\pi_k \in [0,1]}{\text{argmax}}\, \mathcal{C}\left((1-\pi_k)F_{k-1} + \pi_k f_k^*\right). \tag{12}$$

There is an alternative method called partial updating in which each new component and its weight are estimated at the same time, which is preferable since it may result in more robust and reliable estimation:

$$\left\{f_k^*, \pi_k^*\right\} = \underset{\pi_k, f_k}{\text{argmax}}\, \mathcal{C}\left((1-\pi_k)F_{k-1} + \pi_k f_k\right). \tag{13}$$

All the equations for updating weight assigned to feature vector X_t, mixture weight value, and mean and covariance matrix are estimated as follows [14]:

$$\begin{aligned} w^n(X_t) &= \frac{f_k\left(X_t \mid \Phi_k^{(n)}\right)}{\pi_k^n f_k\left(X_t \mid \Phi_k^{(n)}\right) + \left(1-\pi_k^n\right)F_{k-1}(X_t \mid \Psi_{k-1})}, \\ \gamma_t\left(\Phi_k^{(n)}\right) &= \frac{w^n(X_t)}{\sum_{t=1}^{T_j} w^n(X_t)}, \\ {\pi_k}^{n+1} &= \frac{1}{T}\sum_{t=1}^{T_j}\pi_k^n w^n(X_t), \\ {\mu_k}^{n+1} &= \sum_{t=1}^{T_j}\gamma_t\left(\Phi_k^{(n)}\right)\cdot X_t, \\ \Sigma_k^{n+1} &= \sum_{t=1}^{T_j}\gamma_t\left(\Phi_k^{(n)}\right)\cdot\left(X_t - \mu_k^{n+1}\right) \\ &\quad \times \left(X_t - \mu_k^{n+1}\right)^{Tr}. \end{aligned} \tag{14}$$

Moreover, in order to speed the converging process up and find the minimum number of Gaussian component in the final mixture, current mixture model F_k should be updated globally over training data samples before adding the next component. For example, in the GMM with k components, denoted by F_k, the kth component can be reestimated for $k = 1, \ldots, K$ when the reminder of the mixture model is assumed to be fixed. This procedure continues iteratively until the objective function value reaches a local maximum. It means that after obtaining a mixture model F_K, we could update each component f_k and its weight over all training feature vectors by using the same updating equations.

3.7. Criterion for Model Selection. The process of adding new mixture component to previous mixture model continued incrementally and recursively until the optimal number of mixtures is met. The set of Gaussian components selected should represent the space covered by the feature vectors. For this purpose the selected strategy to stop the adding process is a criterion-based one called Bayesian inference criterion (BIC). It can be represented as the following [15]:

$$\text{BIC}(k) = -\mathcal{C}(F_k) + M_k \log\left(T_j\right), \tag{15}$$

where $\mathcal{C}(F_k)$ is the log-likelihood function of the mixture model over all training data, M_k is the number of parameters used in model F_k, and T_j denotes total number of training

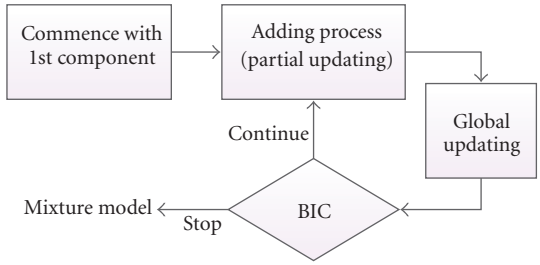

Figure 3: Block diagram of learning GMM using adapted BML technique.

data for the jth pathology class. The second term in BIC equation is a penalty term for the number of parameters in the model. BIC is closely related to Akaike information criterion (AIC) [38, 39] but the penalty term is larger in BIC than in AIC. Figure 3 shows a brief review of all mentioned processes to train a GMM for each available pathological condition in order.

3.8. Decision Rule. The likelihood of a feature vector X given a Gaussian model $\mathcal{L}$ is defined as

$$\mathcal{L}(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X \mid \mu_k, \Sigma_k), \tag{16}$$

where μ_k and Σ_k are mean and covariance parameters for a set of K Gaussians and π_k is a normalizing factor that also weights them appropriately and constrained such that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Each of the trained mixture models $\mathcal{L}_j$ approximates the distribution of the features of one class only. There is significant structural difference between cry signals of infants with different pathologies. Therefore we can assume that the distributions of the features of each cry signal are different. As a result, when classifying a feature vector X_i belonging to pathology class i we will expect $\mathcal{L}_i(X_i) > \mathcal{L}_j(X_i)$, $\forall j \neq i$. One of the common decision rules is to select the hypothesis that has the highest likelihood value called the maximum likelihood (ML) decision criterion:

$$\text{Pathology Class \#} = \max_j \left\{ \mathcal{L}_j(X) \right\}. \tag{17}$$

Likelihood values of undertest infant's cry signal were computed according to generated Gaussian mixture model for each Class$_j$, $j = 1, \ldots, C$, then by the use of the ML rule the decision was made. Figure 4 shows how mixture models trained on the MFCCs are associated with different pathological conditions. NHF and NHP are the total number of Gaussian components which can describe models of healthy full-term and premature infants, respectively. Similarly, NSF$_i$ and NSP$_i$ are the numbers of Gaussian components for full-term and premature infants with pathology i, respectively.

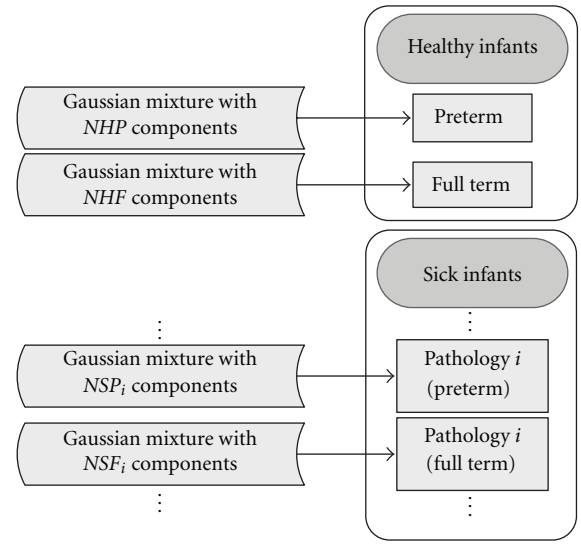

Figure 4: Mapping of estimated GMMs to pathological conditions.

4. Experiments

For implementation the cry database was split into two disjoint subsets for training and testing. Almost 63% of total cry signals were utilized for the training phase and the rest for testing phase. After creating GMMs for each available pathological condition in the cry database using adapted BML method, we can assess what accuracy rate it may have. The MFCC order has also been studied experimentally for speech recognition. A total number of 12 MFCCs are common in speech processing [20, 23] and they are computed directly from the data. It is that same number which is used in [40] to recognize speech emotion via HMMs. The energy within a frame is also an important feature that can be easily obtained. For better performance, the 0th cepstral coefficient $\{c_0\}$ is appended to the feature vector as the 13th feature. The initial coefficient represents the average energy (weighting with a zero-frequency cosine) same as role of the log energy $\{E = \log \sum_{n=1}^{N} S_n^2\}$ [29, 30]. Therefore, used feature vectors are composed of the first 13 MFCCs $\{c_n, n = 0, 1, 2, \ldots, L = 12\}$. As we defined earlier, L is the desired length of cepstrum and it is a fixed parameter. The higher-order MFCC does not further reduce the relative error rate with a typical speech recognition system in comparison with the 13th-order MFCC, which indicates that the first 13 coefficients already contain most salient information needed for speech recognition [29]. We made full use of both initialization methods with decay parameter $p = 0.05$ and linear scaling factor $\beta = -0.5$ values for the parameters to overcome overfitting and the small covariance matrix which was created after several iterations. It is those same parameter values which are set and designed for BML algorithm in large-vocabulary continuous speech recognition tasks in [14]. After that only samples in a subset with equal sample weights were used to estimate the mean and covariance matrix of Gaussian components. The presented method had a vital role to achieve good performance and avoid the overfitting problem in the learning step.

In the first experiment, the NCDS was tested on several multipathology classification tasks. It consisted of all

Table 2: Obtained accuracy rate for multipathology classification task (%).

Frame duration method	20 msec		25 msec		30 msec	
	EM based	ABML	EM based	ABML	EM based	ABML
1	100	100	100	100	100	100
2	100	100	100	100	100	100
3	100	100	0	50	0	50
4	75	100	75	100	75	50
5	100	88.9	100	100	100	100
6	100	100	100	100	100	100
7	80	60	40	60	20	40
8	100	100	100	100	100	100
9	0	0	0	0	0	0

aforementioned conditions in the cry database. It could be difficult to evaluate the effectiveness of the created GMMs for modeling and adapted BML method for learning the mixture model by extracting a single feature from a small cry database. Nevertheless our results show that it had a better accuracy rate compared with conventional EM-based method for GMMs as our reference system. It is worth mentioning that the GMMs created by the EM-based re-estimation method for each class were trained by setting the number of components equal to that of mixture model learned by adapted BML method. Here, we address the question of the classification performance with respect to the frame length by evaluating the system with different frame durations but same overlap percentages (30%) between two consecutive windows. In order to extract MFCCs, frames with different durations are used while 30% overlap was introduced between two consecutive windows. Table 2 shows the obtained accuracy rate for all 9 different groups of infants in the order they are shown in Table 1 for frame durations 20 msec, 25 msec, and 30 msec. It can be seen that both methods delivered great performances for most pathology classes, but the presented method had better final outcomes. For the premature infants with "Coarctation of Aorta" it seemed that the learned pattern was not well defined enough to be capable of accurately classifying them. We believe this was due to either small number of training samples (6 and 4 samples of infants' cry for training and testing, resp.) or used pathologically noninformed features for this disease.

To better understand the effect of frame duration on the performance, the mean values of classification accuracy rates are computed from frequency distribution over 9 pathology classes and they are given by [41]

$$\text{Mean} = \frac{\sum_{i=1}^{9} f_i m_i}{\sum_{i=1}^{9} f_i}, \tag{18}$$

where f_i and m_i show the frequency and obtained accuracy rate for ith pathology class. The results in Figure 5 show that the performances of both EM-based and adaptive BML methods degrade when the frame duration increases. Therefore, the system performance is the best when using 20 msec frame length to extract MFCCs. In addition, Figure 5 says that the presented adaptive BML method, on the average, works better than the EM-based method when frame duration varies.

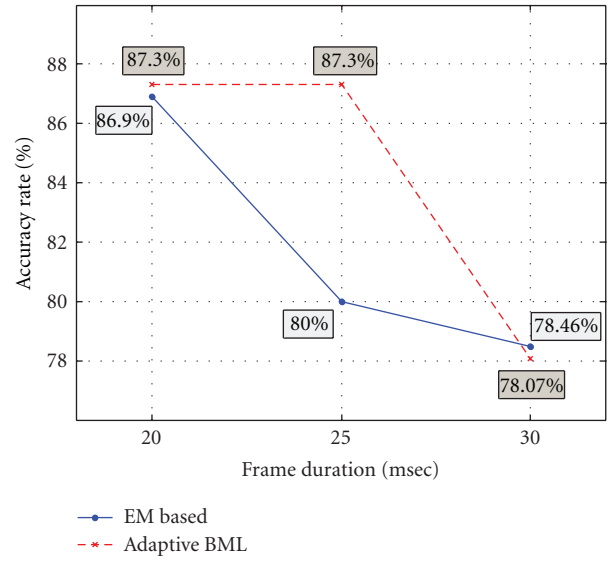

Figure 5: Mean classification accuracy rates.

The coefficient of variation (CV) is particularly useful for representing the reliability of performance tests which is the coefficient of dispersion based on standard deviation. It gives the standard deviation as a percentage of the mean values as follows [41]:

$$\text{Standard Deviation} = \sqrt{\frac{\sum f_i m_i^2 - \left(\sum f_i m_i\right)^2 / \sum f}{\sum f_i - 1}}, \tag{19}$$

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%.$$

In Figure 6 we present this coefficient of dispersion for both techniques with different frame durations. The larger the CV is, the more the performance varies. Since the coefficient of variation is less for shorter segments, their performances are therefore more consistent. The CV values for adaptive BML method are much less than those of EM-based method for frame length 25–30 msec, although they are so close to each other for frame length 20 msec.

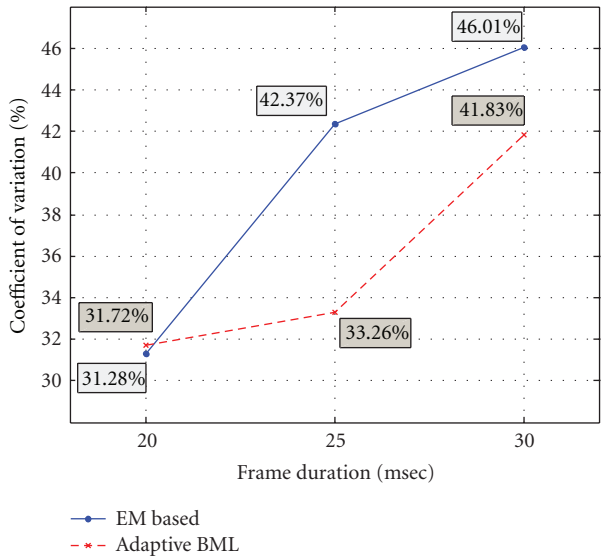

Figure 6: Coefficient of variation.

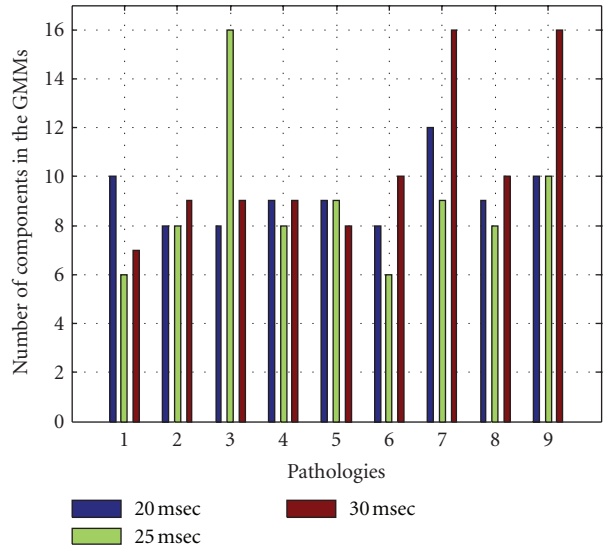

Figure 7: Number of components.

The large cry-signal database is required to make well-defined mixture model to keep the number of components as small as possible. As we said earlier, in large cry-signal samples, a small number of outliers are to be expected. The number of components that minimize BIC in each mixture models is shown in Figure 7.

The second experiment was designed to test the diagnostic system for binary classification task in which all cry data in database were organized into two separate groups, namely, healthy and pathological. Note that, here the frame length of 30 msec with 30% overlap has been used. In this experiment for each defined infant's class, a GMM which fitted extracted data from cry signals is trained by utilizing adapted BML technique. The number of components in the mixture models for healthy and pathological classes was estimated 9 and 12, respectively. These numbers of components were employed in learning steps of mixture models by conventional EM-based method just like in the previous experiment. Table 3 displays two confusion matrices that allow visualization of the performance of each method in the binary classification problem. These two matrices, for (a) the proposed method and (b) conventional EM-based method, are provided for comparison which makes it easy to see if the system is confusing two predefined classes by containing the number of healthy and pathological infants that were correctly classified or mistakenly classified.

True positive (TP) and true negative (TN) rates are defined for a two by two confusion matrix, as calculated using the following equations [42]:

$$\mathrm{TP} = \frac{d}{c+d}, \tag{20}$$

$$\mathrm{TN} = \frac{a}{a+b}, \tag{21}$$

where "a" and "d" are the numbers of correct predictions that an instance is negative or positive, respectively. The other parameters, "b" and "c" are defined in a similar way by counting the number of incorrect predictions that an instance is positive or negative accordingly. In the pathology diagnostics system, true positive rate or test sensitivity shows the ability of the system to correctly detect infants with disease, whereas true negative rate or test specificity demonstrates the ability of the system to correctly identify those without disease. Both statistical measures of the two methods are summarized in Table 4.

The ramification of false positive (type I error) and false negative (type II error) in some cases especially medical examinations are not the same [18]. One can be more costly and irrecoverable than the reverse situation. Preferably, the classifier should be able to provide a balanced degree of

Table 3: Confusion matrix for defined binary classification task.

(a) Proposed ABML-GMM method

	Predicted classes	
Actual classes	Pathological	Healthy
Pathological	21	5
Healthy	3	20

(b) Conventional EM-GMM method

	Predicted classes	
Actual classes	Pathological	Healthy
Pathological	26	0
Healthy	6	17

Table 4: Obtained two statistical measures for binary-classification problem.

	EM-GMM	ABML-GMM
Test sensitivity	100%	80.77%
Test specificity	73.91%	86.96%

predictive accuracy (ideally 100%) for both the minority and majority classes in our imbalanced data which is a direct result of the nature of the cry database.

5. Conclusion

A newborn cry-based diagnostic system (NCDS) based on extracting mel-frequency cepstral coefficients (MFCCs) from infant's cry signal is presented in this paper. For all cry samples which belong to the healthy infant class or pathological infant classes with different physiology conditions, a mixture model with separate Gaussian pool was estimated as a cry pattern. Adapted boosting mixture learning (BML) method was introduced to train mixture models. Some advanced techniques of signal processing and machine learning were employed in different parts of the learning process such as adding new component, weighting function for samples, model selection, and global re-estimation of parameters. In multi-pathological classification tasks, results show that, on the average, the presented method achieved a higher classification accuracy rate to identify infants' diseases than EM-based re-estimation algorithm for Gaussian mixture models (GMMs). The performance and reliability of adaptive BML-GMMs are the best when using 20 msec as a frame duration and gradually degrade when the length increases further. The results demonstrate that the adaptive BML method can provide better classification accuracy rate than EM-based method with higher system reliability. For binary classification problem, with 30 msec frame duration (the worst-case scenario), adapted BML can identify full-term and premature healthy infants better than EM-based method, but on the other hand it delivers a little lower performance than EM-based method for sick infants. However, adapted BML provides better balanced degree of predictive accuracy for both the minority and majority classes (test sensitivity 80.77% and test specificity 86.96%).

Acknowledgments

The authors would like to thank Dr. Barrington and the members of Neonatology Group of Mother and Child University Hospital Center in Montreal (QC) for their dedication in the collection of the infant's cry database. This work has been funded by a grant from Bill & Melinda Gates Foundation through the Grand Challenges Explorations Initiative.

References

[1] O. Wasz-Hockert, K. Michelsson, and J. Lind, *Twenty-Five Years of Scandinavian Cry Research*, New York, NY, USA, 1985.

[2] J. Benson, *Social and Emotional Development in Infancy and Early Childhood*, Elsevier, 2009.

[3] O. Wasz-Hockert, *The Infant Cry: A Spectrographic and Auditory Analysis*, Lippincott, Philadelphia, Pa, USA, 1968.

[4] M. J. Corwin, B. M. Lester, and H. L. Golub, "The infant cry: what can it tell us?" *Current Problems in Pediatric and Adolescent Health Care*, vol. 26, no. 9, pp. 325–333, 1996.

[5] O. Galaviz and C. García, "Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system," in *Proceedings of the Advances in Artificial Intelligence (MICAI '05)*, A. Gelbukh and H. Terashima, Eds., vol. 3789, pp. 949–958, Springer.

[6] S. Cano, I. Suaste, D. Escobedo, C. A. Reyes-García, and T. Ekkel, "A combined classifier of cry units with new acoustic attributes," *Progress in Pattern Recognition, Image Analysis and Applications*, vol. 4225, Springer, 2006.

[7] E. Amaro-Camargo and C. Reyes-García, "Applying statistical vectors of acoustic characteristics for the automatic classification of infant cry," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, D. S. Huang, L. Heutte, and M. Loog, Eds., vol. 4681, pp. 1078–1085, Springer, Berlin, Germany, 2007.

[8] J. Orozco and C. A. R. Garcia, "Detecting pathologies from infant cry applying scaled conjugate gradient neural networks," in *Presented at the European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2003.

[9] K. Lind and K. Wermke, "Development of the vocal fundamental frequency of spontaneous cries during the first 3 months," *International Journal of Pediatric Otorhinolaryngology*, vol. 64, no. 2, pp. 97–104, 2002.

[10] M. Hariharan, S. Yaacob, and S. A. Awang, "Pathological infant cry analysis using wavelet packet transform and probabilistic neural network," *Expert Systems With Applications*, vol. 38, no. 12, pp. 15377–15382, 2011.

[11] S. Cano, I. Suaste, D. Escobedo, C. A. Reyes-García, and T. Ekkel, "A radial basis function network oriented for infant cry classification," in *Progress in Pattern Recognition, Image Analysis and Applications*, J. Martínez-Trinidad, J. A. C. Ochoa, and J. Kittler, Eds., vol. 3287, pp. 15–36, Springer, Berlin, Germany, 2004.

[12] E. Bauer and R. Kohavi, "Empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1, pp. 105–139, 1999.

[13] A. Divakaran, *Multimedia Content Analysis: Theory and Applications (Signals and Communication Technology)*, Springer, 2009.

[14] D. Jun, Y. Hu, and H. Jiang, "Boosted mixture learning of gaussian mixture hidden markov models based on maximum likelihood for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2091–2100, 2011.

[15] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[16] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[17] D. O'Shaughnessy, "Invited paper: automatic speech recognition: history, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[18] H. He, *Self-Adaptive Systems for Machine Intelligence*, 2011.

[19] W. Holmes, *Speech Synthesis and Recognition*, Taylor & Francis, 2nd edition, 2001.

[20] J. J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*, Prentice Hall, PTR, 1993.

[21] A. H. Gray Jr. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 3, pp. 207–217, 1974.

[22] J. Makhoul and R. Viswanathan, "Adaptive preprocessing for linear predictive speech compression systems," *Journal of the Acoustical Society of America*, vol. 55, pp. 475–476, 1974.

[23] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, (For HTK Version 3. 4)*, Cambridge University Engineering Department, 2006.

[24] I. Mporas, T. Ganchev, O. Kocsis, and N. Fakotakis, "Context-adaptive pre-processing scheme for robust speech recognition in fast-varying noise environment," *Signal Processing*, vol. 91, no. 8, pp. 2101–2111, 2011.

[25] O. O. Akande and P. J. Murphy, "Estimation of the vocal tract transfer function with application to glottal wave analysis," *Speech Communication*, vol. 46, no. 1, pp. 15–36, 2005.

[26] R. Flynn and E. Jones, "Combined speech enhancement and auditory modelling for robust distributed speech recognition," *Speech Communication*, vol. 50, no. 10, pp. 797–809, 2008.

[27] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[28] M. Benzeghiba, R. De Mori, O. Deroo et al., "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[29] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing : A Guide To Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[30] D. O'Shaughnessy, *Speech Communications: Human and Machine*, Wiley-IEEE Press, Montréal, Canada, 2000.

[31] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[32] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the Em algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[33] S.-K. Ng and G. J. McLachlan, "Speeding up the EM algorithm for mixture model-based segmentation of magnetic resonance images," *Pattern Recognition*, vol. 37, no. 8, pp. 1573–1589, 2004.

[34] A. Berlinet and C. Roland, "Acceleration schemes with application to the EM algorithm," *Computational Statistics and Data Analysis*, vol. 51, no. 8, pp. 3689–3702, 2007.

[35] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[36] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[37] S. Rosset, "Robust boosting and its relation to bagging," in *Presented at the Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Ill, USA, 2005.

[38] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[39] T. Bengtsson and J. E. Cavanaugh, "An improved Akaike information criterion for state-space model selection," *Computational Statistics and Data Analysis*, vol. 50, no. 10, pp. 2635–2654, 2006.

[40] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[41] D. Zill and W. S. Warren, *Advanced Engineering Mathematics*, 4th edition, 2011.

[42] R. Kohavi and F. Provost, "Glossary of terms," *Editorial For the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, vol. 30, pp. 271–274, 1998.