Naithani *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2018) 2018:1
DOI 10.1186/s13636-018-0124-x

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Automatic segmentation of infant cry signals using hidden Markov models

Gaurav Naithani[1*†], Jaana Kivinummi[2†], Tuomas Virtanen[1], Outi Tammela[3], Mikko J. Peltola[4] and Jukka M. Leppänen[2]

Abstract

Automatic extraction of acoustic regions of interest from recordings captured in realistic clinical environments is a necessary preprocessing step in any cry analysis system. In this study, we propose a hidden Markov model (HMM) based audio segmentation method to identify the relevant acoustic parts of the cry signal (i.e., expiratory and inspiratory phases) from recordings made in natural environments with various interfering acoustic sources. We examine and optimize the performance of the system by using different audio features and HMM topologies. In particular, we propose using fundamental frequency and aperiodicity features. We also propose a method for adapting the segmentation system trained on acoustic material captured in a particular acoustic environment to a different acoustic environment by using feature normalization and semi-supervised learning (SSL). The performance of the system was evaluated by analyzing a total of 3 h and 10 min of audio material from 109 infants, captured in a variety of recording conditions in hospital wards and clinics. The proposed system yields frame-based accuracy up to 89.2%. We conclude that the proposed system offers a solution for automated segmentation of cry signals in cry analysis applications.

Keywords: Infant cry analysis, Acoustic analysis, Audio segmentation, Hidden Markov models, Model adaptation

1 Introduction

For several decades, there has been an ongoing interest in the connection of acoustic characteristics of infant cry vocalizations with infant health and developmental issues [1–4]. Atypicalities in specific features of cry (e.g., fundamental frequency) have been linked with diagnosed conditions such as autism, developmental delays, and chromosome abnormalities [5, 6] and with risk factors such as prematurity and prenatal drug exposure [7, 8]. These findings have generated hope that cry analysis may offer a cost-effective [9], low-risk, and non-invasive [10, 11] diagnostic technique for early identification of children with developmental and health problems. The need for detecting health problems and risks (e.g., as pointed out by [6]) as early as possible is important because the plasticity of the developing brain and the sensitive periods of skill formation at the very early age offer the best chances to support optimal development by rehabilitation and medical care [12–16].

An infant cry signal consists of a series of expirations and inspirations separated by bouts of silence. These will be referred to as *expiratory* and *inspiratory phases* in this paper. A cry signal captured in a realistic environment (e.g., pediatric ward of a hospital) may contain extraneous sounds (e.g., non-cry vocals produced by the infant, vocals of other people present in the room, and background noise contributed by the surrounding environment or by the recording equipment itself). A cry signal recording can thus be thought of being composed of what we call the regions of interest, namely, expiratory and inspiratory phases, and extraneous regions consisting the rest of the audio activity contained in the recording, termed as *residual* in this paper. Figure 1 is an example of a chunk of a cry recording captured in hospital environment.

In realistic clinical situations, the recordings are affected by the acoustic environment including the room acoustics and other sound sources present during the recording. The cry signal itself is affected by several factors related to

*Correspondence: gaurav.naithani333@gmail.com
†Equal contributors
[1]Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 10, Tampere, Finland
Full list of author information is available at the end of the article

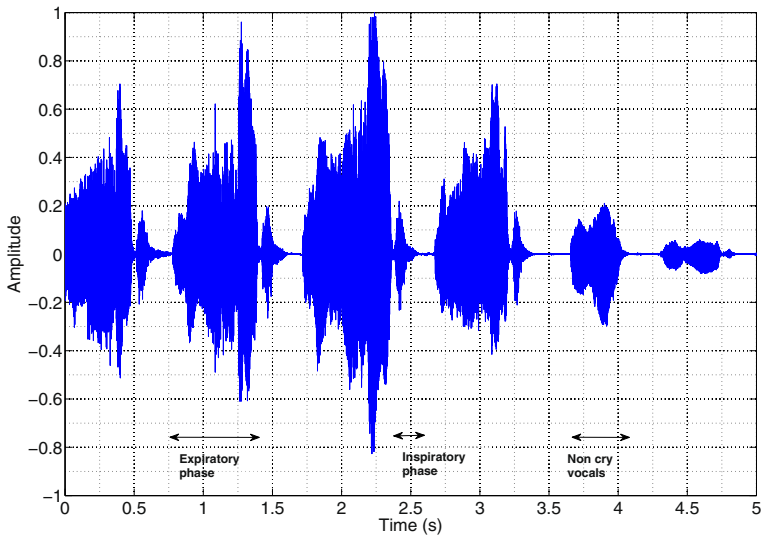

Fig. 1 An example of a chunk of infant cry signal showing expiratory and inspiratory phases and non-cry vocals present in the recording (categorized as residual)

the state of infants' health and development [6], age [17], size [18], reason of crying [19], and arousal [20].

In a cry analysis system meant to work with recordings captured in realistic environments, there is often a need for a pre-processing system which is able to differentiate the regions of interest (i.e., expiratory and inspiratory phases) from extraneous acoustic regions (i.e., residual). The need for identifying the expiratory and inspiratory phases as separate classes arises from the fact that they differ in their properties, e.g., fundamental frequency, harmonicity, and time duration. Successful extraction of these are of significant interest when the system output is used as a diagnostic tool where the relation of these properties to infant neuro-developmental outcomes can be explored. Manual annotation of cry recordings is prone to errors and is rendered unfeasible when the number of recordings to be annotated is large. The segmentation mechanism in any such cry analysis system thus needs to be automatic and should be able to work with material captured in diverse recording contexts.

Various methods have been previously used in the field of infant cry analysis to deal with the problem of identifying the useful acoustic regions from cry recordings, for example, manual selection of voiced part of recordings [4, 21], cepstral analysis for voicing determination [22], harmonic product spectrum (HPS) based methods [23], short-term energy (STE) histogram based methods [24, 25], and *k*-nearest neighbor algorithm based detection [26]. Most of these methods have treated inspiratory phases as noise, and primary attention has been focused on extraction of expiratory phases. The relevance of anatomical and physiological bases of inspiratory phonation has been pointed out by Grau et al. [27]. Previously, Aucouturier et al. [28] have used the hidden Markov model (HMM) based method for automatic segmentation of cry signals while treating inspiratory vocalization as a separate class. They utilized standard mel-frequency cepstral coefficients (MFCC) as audio features and employed HMM topology consisting of a single state for each target class. Abou-Abbas et.al [29] proposed a similar HMM based method utilizing delta and delta-delta features along with MFCCs and experimenting with more number of HMM states for each class. Similarly, Abou-Abbas et.al. [30] proposed cry segmentation using different signal decomposition techniques. Hidden Markov models for cry classification instead of detection have been studied by Lederman et.al. [31, 32].

In this paper, we propose an HMM based method for identifying useful acoustic regions from cry recordings captured under diverse recording conditions. The diversity of recording conditions includes acoustic conditions of recording, types of cry trigger, and infant-related factors which are known to affect acoustic characteristics of cry. Sections 4.1 and 4.2 describe this in detail. The work presented here distinguishes itself from similar previous efforts by proposing the use of fundamental frequency and aperiodicity (see Section 2.1) as audio features in addition to conventionally used features, e.g., MFCCs and their first and second order derivatives. We show that this yields an improvement in segmentation performance. Moreover, we show that the proposed system is able to adapt to material recorded in unseen acoustic environments for which

it has not been trained. We use a combination of feature normalization and semi-supervised learning for this adaptation problem.

The paper follows the following structure. Section 2 explains the implementation of the proposed system, Section 3 explains the model adaptation techniques, Section 4 describes the data used in experiments, Section 5 describes the evaluation and presents the obtained results, and, finally, Section 6 provides some concluding remarks with suggestions for future directions of this work.

2 Proposed method

In order to analyze infant cry recordings captured in realistic environments containing interfering sources, the goal is to segment cry recordings into three classes, namely, expiratory phases, inspiratory phases, and residual. The residual class consists of all acoustic regions in the cry recording except the ones covered by the other two classes. A supervised pattern recognizer based on hidden Markov models (HMM) with Gaussian mixture model (GMM) densities [33] is used for segmentation. An HMM is a statistical model which models a generative time sequence characterized by an underlying hidden stochastic process generating an observable sequence [34]. HMMs have been widely used in automatic speech recognition (e.g., [35]) to model variability in speech caused by different speakers, speaking styles, vocabularies, and environments.

Figure 2 depicts the block diagram of the segmentation process. Each cry recording under investigation is divided into windowed overlapping short time frames. For each such frame, the HMM pattern recognizer outputs a set of observation probabilities of the three classes being active in that frame. These probabilities are decoded using the Viterbi algorithm. Decoding here refers to the process of finding the best path in the search space of underlying HMM states that gives maximum likelihood of the acoustic feature vectors from the cry signal under investigation. It outputs a class label for each frame of the signal, and this information is used to identify the regions of interest (i.e., expiratory and inspiratory phases) in the cry signal. The overall implementation can thus be described in three stages, namely, feature extraction, HMM training, and Viterbi decoding. These stages are described in the following subsections.

2.1 Feature extraction

Mel-frequency cepstral coefficients (MFCC) are used as the primary audio features. They have been widely used in audio signal processing problems, for example, speech recognition [36], audio retrieval [37], and emotion recognition in speech [38]. Frame duration of 25 ms with 50% overlap between consecutive frames and Hamming window function was used for extracting the MFCCs. For each frame of the signal, a 13-dimensional MFCC feature vector, $\mathbf{x} = [x_1, x_2 ..... , x_{13}]^T$, is extracted, which includes the zeroth MFCC coefficient; here, T represents the matrix transpose. The sampling frequency for each audio signal is 48 kHz. In conjunction with MFCCs, the following additional features are investigated.

1. *Deltas and delta-deltas*: MFCCs are static features and provide a compact spectral representation of only the corresponding frame. Temporal evolution of these features might be useful for segmentation purposes since HMMs assume each frame to be conditionally independent of the previous ones given the present state. This temporal dynamics is captured by computing the time derivatives of MFCCs, known as delta features. Similarly, temporal dynamics of delta features can be captured by computing their time derivatives, known as delta-delta features. For 13 MFCCs per frame, we have 13 delta coefficients and 13 delta-delta coefficients. The use of these time derivatives also means that the system is non-causal.
2. *Fundamental frequency* ($F0$): Inspiratory phases are known to have higher fundamental frequency ($F0$) than expiratory phases [27]. This property can be exploited for segmentation purposes by including $F0$

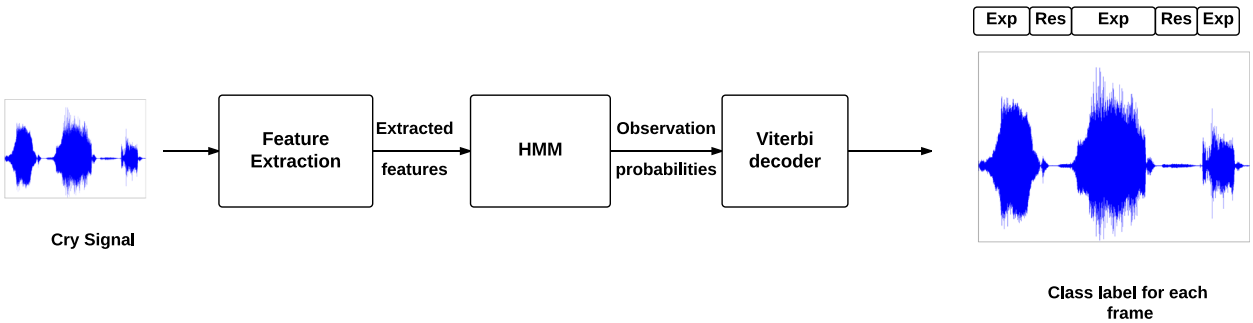

Fig. 2 Block diagram of the audio segmentation system

as an audio feature. The YIN algorithm [39] is a popular pitch estimation algorithm, which has been found to perform well in the contexts of speech [40] and music [41] signals. A freely available MATLAB implementation of the algorithm is used in the proposed system [42], and one $F0$ value is obtained for each frame. We found YIN algorithm to be suitable for this dataset as $F0$ values were empirically found to be between 200 and 800 Hz and very few instances of hyperphonation ($F0 < 1000$ Hz) [6] were observed.

3. *Aperiodicity* : Aperiodicity in this study refers to the proportion of aperiodic power in the signal frame and is computed through the YIN algorithm. In order to compute an $F0$ estimate, the YIN algorithm employs a function known as *cumulative mean normalized difference function*. The minima of this function that subscribes to certain conditions gives an estimate of the fundamental period of the signal frame. The value of the function at this minima is proportional to the aperiodic power contained in the signal frame. A detailed mathematical treatment can be found from the original paper [39]. One aperiodicity value is obtained corresponding to each frame.

2.2 Cry modeling using HMMs

The available dataset is manually annotated and divided into training and test sets, as will be described in detail in Section 4. Features extracted from all audio files in the training dataset for a particular target class are concatenated to give training feature matrix $\mathbf{X}_i$, i being the index of the target class. Using these feature matrices, three separate HMM models are trained corresponding to the three target classes: expiratory phases, inspiratory phases, and residual. The probability density function (pdf) of each HMM state is modeled with Gaussian mixture models (GMMs).

Training involves estimating HMM parameters, λ_i (i.e, weight, mean, and covariance of component Gaussians and state transition probabilities), which best fits the training data $\mathbf{X}_i$. Probabilistically, it is framed as problem of maximizing probability of an HMM model given the training data $\mathbf{X}_i$, which in turn can be framed as maximum likelihood estimation problem, i.e.,

$$\lambda_i^{\text{opt}} = \arg\max_{\lambda} P(\mathbf{X}_i|\lambda_i) \quad (1)$$

where λ_i^{opt} indicates the optimal model for i^{th} class. For this, the standard Baum-Welch algorithm [43], an expectation maximization algorithm used to estimate HMM parameters, is used. AHTO toolbox of the Audio Research Group, Tampere University of Technology, is used for this purpose. Fully connected HMMs are used with each state having equal initial state probability. It also means all entries of initial state transition probability matrix are non-zero and equal. For state means and covariances, k-means clustering initialization is used.

Two parameters have to be chosen for each HMM: S, the number of states used to adequately model the class, and C, the number of Gaussian components in the corresponding GMM used to model each state of the HMM. The effect of both these parameters on system performance has been investigated and will be discussed in Section 5. The number of states and component Gaussians in the three HMMs are denoted by S_{exp} and C_{exp}, S_{ins} and C_{ins}, and S_{res} and C_{res} for expiratory phase, inspiratory phase, and residual, respectively. HMMs trained for the three target classes are then combined to form a single HMM having a combined state space and transition probability matrix. State transitions from any state of one model to any state of another model are possible, in other words, the combined HMM model is fully connected. The combined model has a transition probability matrix having dimensions $(S_{\text{exp}} + S_{\text{ins}} + S_{\text{res}}) \times (S_{\text{exp}} + S_{\text{ins}} + S_{\text{res}})$. The probability of transition from one model to another depends upon model priors and *inter-model transition penalty*, a parameter similar to HTK toolkit's [44] *word transition penalty* parameter. *Inter-model transition penalty* penalizes model transition from one model to another and has to be empirically determined (we have used a value of -1 in this paper). The model priors are calculated simply by counting the occurrences of the corresponding class from the annotated data.

HMM parameters of this combined model are used for Viterbi decoding of observation probability outputs in the following section. Figure 3 depicts the combined HMM.

2.3 Viterbi decoding

Features extracted from the cry recording to be segmented are fed to the three HMM models, each trained for a particular target class. For each frame of the recording,

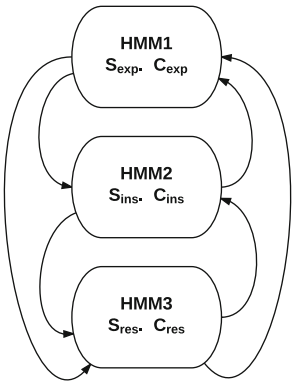

Fig. 3 HMM models for individual classes are combined into a single fully connected HMM

the HMM outputs the probabilities of its constituent states being active in that frame. Three observation probability matrices are generated corresponding to three HMMs which are combined into a single matrix $\mathbf{O}_{\text{comb}}$ as depicted in Fig. 4. The Viterbi algorithm is employed upon this combined observation probability matrix using the parameters learned for combined HMM model in the previous section. The algorithm maximizes the probability of occurrence of state sequence $\mathbf{q}$ given a learned HMM λ_{comb} and observation probability matrix $\mathbf{O}_{\text{comb}}$, i.e.,

$$\mathbf{q}^{opt} = \arg\max_{\mathbf{q}} P(\mathbf{q}|\mathbf{O}_{\text{comb}}, \lambda_{\text{comb}}) \tag{2}$$

where $\mathbf{q}^{opt}$ is the state sequence giving maximum likelihood through the combined HMM state space which now consists of $(S_{\text{exp}} + S_{\text{ins}} + S_{\text{res}})$ states. This output sequence consists of a state assignment for each frame of the recording, which can further be used to give corresponding class assignment for each frame. It is done by identifying the contributing HMM corresponding to the chosen state for that frame. AHTO toolbox is used for Viterbi decoding. Figure 4 shows the implementation of the audio segmentation system. The segmentation results for a 5-s chunk of a cry signal is depicted in Fig. 5.

3 Model adaptation

The proposed audio segmentation system is trained on a dataset recorded in a particular acoustic environment and may not necessarily be able to generalize for a dataset captured in a different acoustic environment. In this section, we will show that the proposed system can be made to work for data recorded in unseen acoustic environments as well. We will train our system on data recorded in a known acoustic environment and use it predict class labels on data recorded in an unseen acoustic environment. Our proposed solution consists of two stages: feature normalization [45] and semi-supervised learning. These will be described in detail in the following subsections.

3.1 Feature normalization

Features extracted from an audio file are normalized by subtracting the mean and dividing it by the standard deviation before feeding it to the HMM. The mean and standard deviation vectors are derived for each audio file separately. This is repeated for each audio file present in the training data (from known environment) as well as the test data (from unknown environment). For a feature vector $\mathbf{F}_{jn}$ extracted from j^{th} frame of n^{th} audio file, the normalized feature vector is given by

$$\mathbf{F}_{jn}^{std} = \frac{\mathbf{F}_{jn} - \boldsymbol{\mu}_n}{\boldsymbol{\sigma}_n} \tag{3}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\sigma}_n$ are mean vector and standard deviation vector, respectively, derived for n^{th} audio file. The divide operation here is element-wise.

3.2 Semi-supervised learning

A semi-supervised learning (SSL) method, known as self training [46], is used to further adapt the HMM models to an unseen acoustic environment. In a classical SSL problem, we have two datasets: labeled data (from a known acoustic environment) and unlabeled data (from a new acoustic environment). The idea behind this method is to generate additional labeled training data using the unlabeled data comprised of audio files recorded in the unseen acoustic environment. The output labels generated by the model for the unlabeled data are treated as true labels, and the models are retrained using the combination of original training data and this newly generated labeled data. Figure 6 depicts this process.

Alternatively, instead of using the entire unlabeled data, a selection of only those frames can be made for which we are confident of the assigned label being true. The likelihoods outputted by HMMs corresponding to three target classes may be used to devise a confidence criterion. In Fig. 4, we have three likelihood matrices corresponding to each target class for each test file. The maximum likelihood for each column of the three matrices is calculated. The ratio between the maximum and second largest value roughly represents how confident we can be about the classification result for a particular frame. We will refer it as the confidence threshold. Only those frames for which this ratio exceeds a certain threshold are chosen. A confidence threshold of 2 was used in this work. Figure 7 shows the procedure of selecting data based on confidence threshold. Data selected this way can be used as additional training data for HMMs corresponding to the three classes.

4 Acoustic material

For this study, we collected cry recordings from two cohorts of infants in Tampere, Finland, and in Cape Town, South Africa. The following subsections describe these two databases and evaluation of the performance of the audio segmentation system on them.

4.1 Database: Tampere cohort

In Tampere, Finland, we captured the recordings at Maternity Ward Units and Neonatal Ward Unit of Tampere University Hospital. The recording period was from April 13 to August 3, 2014. The study followed the stipulated ethical guidelines and was approved by the Ethical Committee of Tampere University Hospital. The cohort consisted of a heterogeneous group of 57 neonates whose chronological ages (i.e., the time elapsed since birth) at recording were from 0 to 5 days as depicted in Table 1. The cohort was not

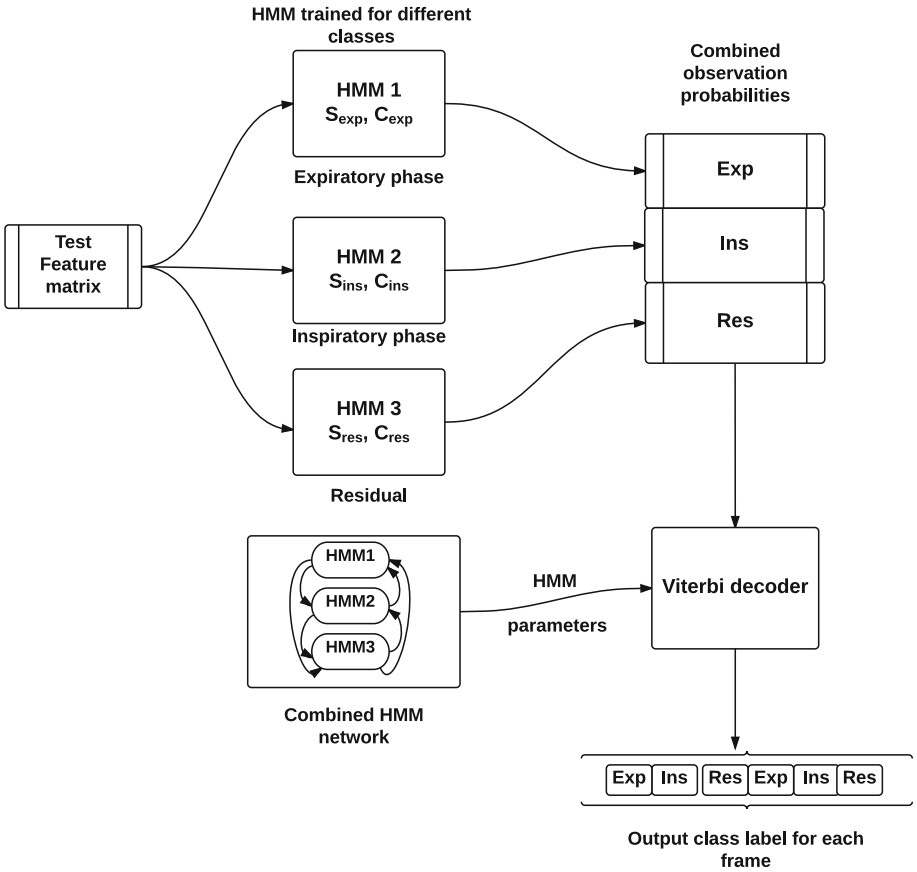

Fig. 4 Implementation of the audio segmentation system. The input to the system is a feature matrix derived from a test audio file. The output is a class label assigned for each frame of the cry signal

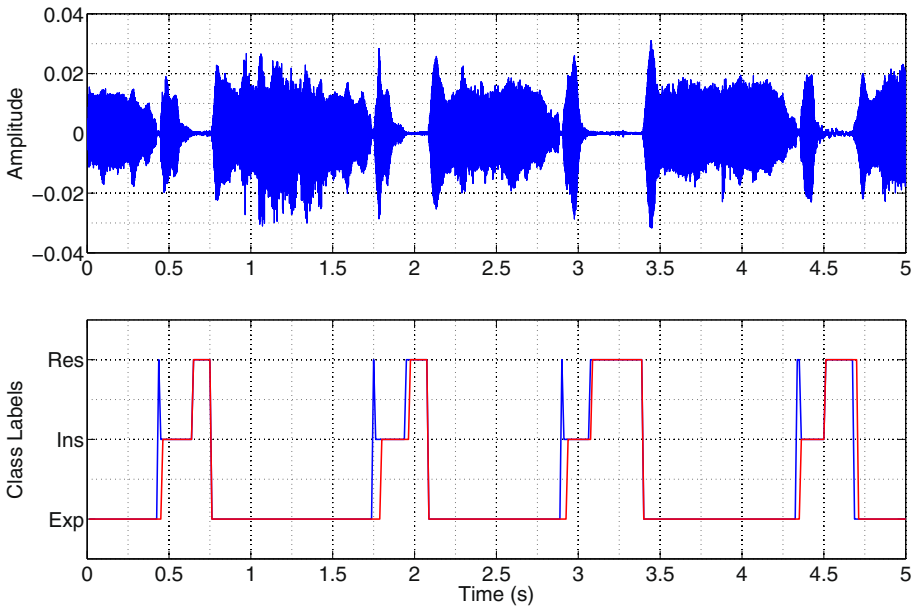

Fig. 5 Audio segmentation results for a chunk of cry signal shown in the top panel. The bottom panel depicts the actual and predicted class labels with blue and red plots, respectively

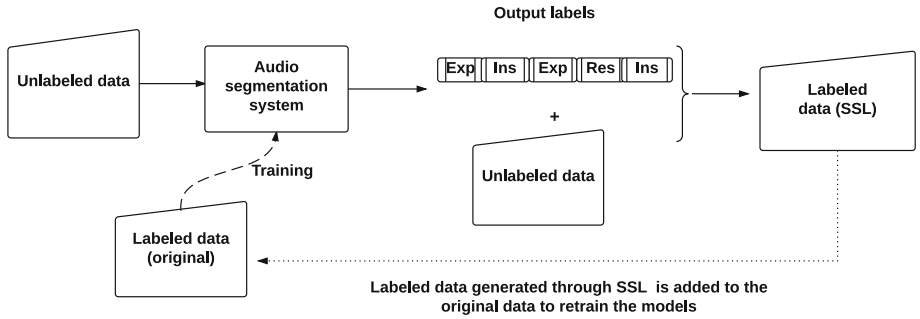

Fig. 6 Semi-supervised learning block diagram

standardized because the target of the present study was to develop a robust tool for identifying infant cry sounds in the captured recordings for general neonate population. In order to minimize the influence of learning and maturation on cry characteristics, the age of the infants was the only standardized variable in the cohort.

The cry samples were captured in a variety of recording conditions. Firstly, the place of recording and the associated acoustic environment varied significantly. It included the hospital corridor, normal pediatric ward, intensive care unit (ICU), waiting room, and nurse's office. Within each room, recordings were captured at different places (e.g., mother's bed, weighing scales, and infant's bed). Secondly, the background sounds present in the recording consisted of human voices (e.g., coughing and speaking) and mechanical sounds (e.g., sound of running water, air conditioning, and diaper tape being opened). Thirdly, infant-related factors (e.g., weight of the infant and prematurity of birth) that are known to influence the acoustic qualities of cry varied . Apart from the recording conditions, the cry-initiating trigger also varied. It included invasive (e.g., venipuncture) and non-invasive (e.g., changing diapers and measuring body temperature) operations, as well as spontaneous cries (e.g., due to hunger or fatigue).

All Tampere recordings were stored as 48 kHz sampling rate, two-channel audio in a 24-bit Waveform audio file (WAV) format. The audio recorder used was Tascam DR-100MK II with RØDE M3 cardioid microphone. For further computation, the mean of the two channels was taken to yield the signal to be segmented. The distance between the infant's mouth and the recorder was kept at

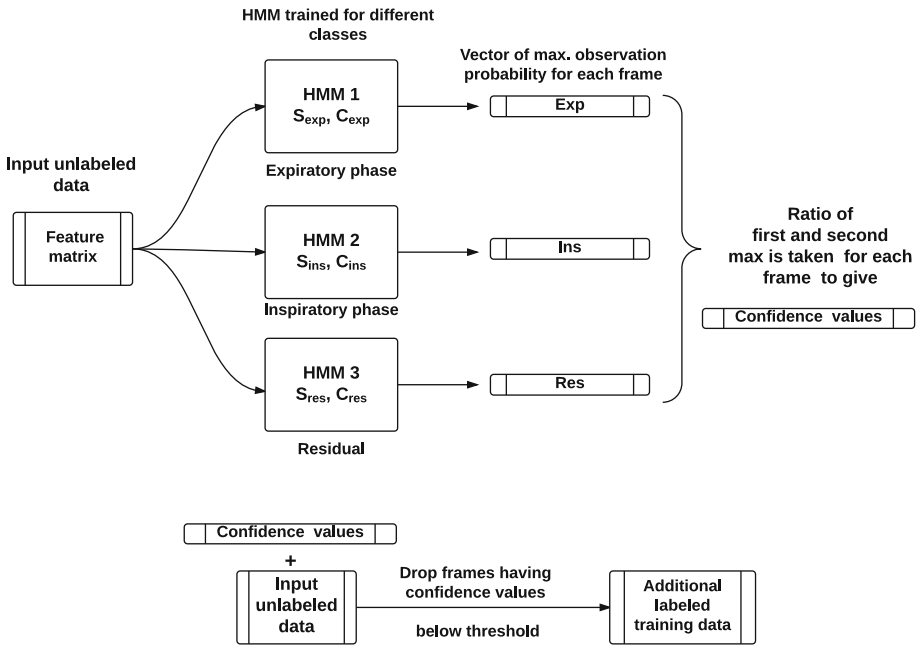

Fig. 7 Selection of data based on confidence threshold for each unlabeled audio file. The model outputs provide the labels for semi-supervised learning

Table 1 The chronological ages of infant subjects in the Tampere cohort

No. of infants	Chronological age (day)
1	0
11	1
29	2
10	3
3	4
2	5
1	Missing info.

approximately 30 cm. Each recording was given a separate number code. The recordings were manually annotated using Audacity [47] application to generate labels for training the HMM models. Figure 8 is a snapshot of the Audacity application showing an example of a chunk of the labeled cry recording.

The database of 57 manually annotated audio recordings spans around 115 min in duration. A total of 1529 expiratory phases were found with a mean duration of 0.95 s and a standard deviation of 0.65 s. Similarly, 1005 inspiratory phases were found with a mean duration of 0.17 s and a standard deviation of 0.06 s. Figure 9 (top) illustrates the distribution of the time durations for expiratory and inspiratory phases for the Tampere cohort.

Note that inspiratory phases were fewer in number and shorter in duration as compared to expiratory phases. Hence, less data were available for training the HMM for inspiratory phases as compared to expiratory phases. Moreover, it needs to be emphasized here that inspiratory phases exhibited more variations throughout the data in comparison to expiratory phases. For example, on the one hand, we had recordings with very short or almost no discernible inspiratory phases, and on the other hand, we had recordings which have unusually prominent inspiratory phases as compared to expiratory phases. It is also possible to observe both these extreme cases within the same recording.

4.2 Database: Cape Town cohort

The other cohort used for this study is being investigated under a larger research project in cooperation with the Department of Psychiatry, University of Stellenbosch, Cape Town. The data were collected in 2014 and consisted of cry recordings of 52 infants whose age was less than 7 weeks (mean 33.5 days, standard deviation 3.5 days). The cry recordings in this database were also manually annotated using the Audacity application. The database of 52 manually annotated audio recordings spans around 75 min in duration. A total of 1307 expiratory phases were found with a mean duration of 1.1 s and a standard deviation of 0.76 s. Similarly, 680 inspiratory phases were found with a mean duration of 0.25 s and a standard deviation of 0.07 s. Figure 9 (bottom) illustrates the distribution of the durations for expiratory and inspiratory phases for the Cape Town cohort.

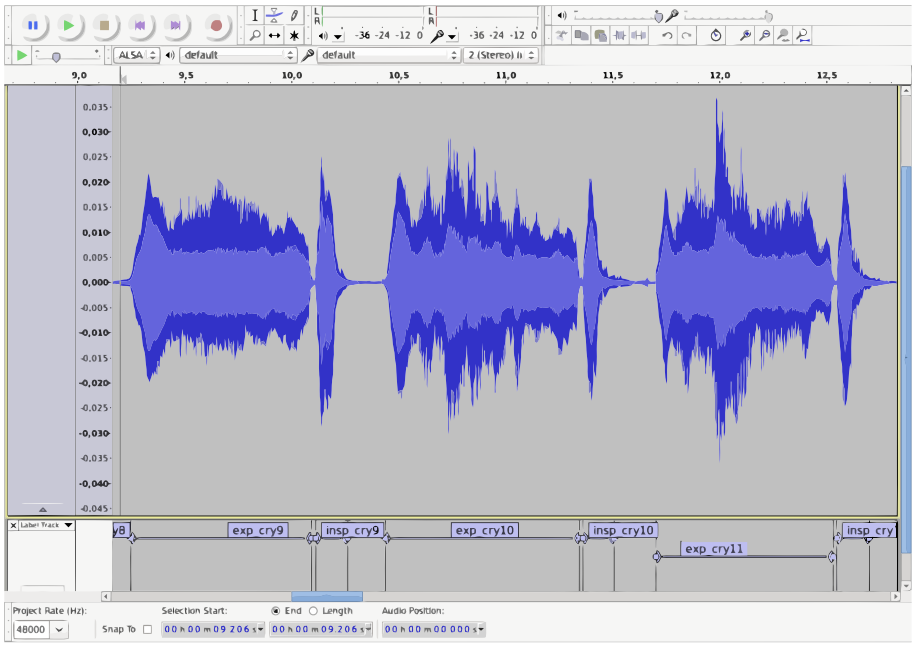

Fig. 8 Snapshot of the Audacity application showing a manually annotated chunk of the cry recording. Expiratory and inspiratory phases are coded by names *exp_cry* and *insp_cry*, respectively

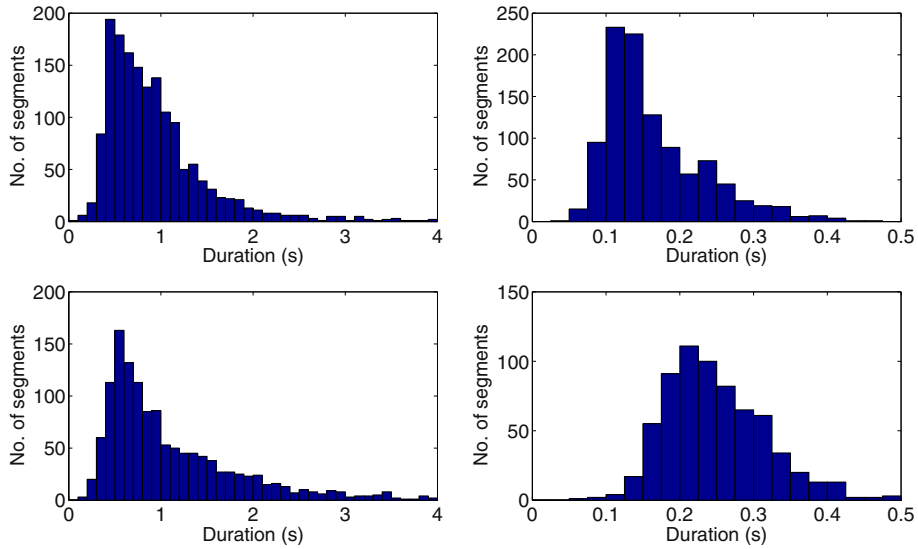

Fig. 9 Distribution of durations of expiratory (left) and inspiratory (right) phases for Tampere (top) and Cape Town (bottom) cohorts

In the Cape Town cohort, the location and procedure of recording were somewhat more standardized than in the Tampere cohort (i.e., the recordings were captured while conducting routine examinations in the same nursing room). The cry trigger used was vaccination (i.e., invasive) or measurement of infant weight at a weighing scale (i.e., non-invasive). All Cape Town recordings were stored as 48-kHz sampling rate, two-channel audio in a 24-bit Waveform audio file (WAV) format. The audio recorder used was Zoom H4n recorder with built-in condenser microphones. The distance between infant's mouth and the recorder was approximately 1.3 m for infants being vaccinated and 70 cm for infants being weighed. Our data collection was conjoined with another study whose protocol required the mic to be a bit far and hence the larger distance between the infant and the recorder as compared to the Tampere cohort. Due to guidelines of the project concerning protection of privacy of the involved participants, we are not able to publish the audio data used in this project.

5 Evaluation

The segmentation performance was evaluated using a five-fold cross-validation framework. In the case of Tampere cohort, the available dataset of 57 cry recordings was divided into five partitions: four partitions of 12 recordings each and one partition of nine recordings. In a similar manner, for the Cape Town cohort, the dataset of 52 cry recordings was divided into five partitions: four partitions of 10 recordings and one partition of 12 recordings. The division was done according to cry codes assigned to the recordings which correspond to the chronological order in which they were captured. In each fold, one of the partitions was used as the test set and the rest of the partitions were used for training. Five such folds were performed with each fold having a different partition as the test set. The output labels generated by the system were compared against the manually annotated ground truth.

For each test file under investigation, the output labels produced by the model were compared against the ground truth (i.e., manual annotations) to calculate the performance metrics. Two metrics have been used in this study to evaluate the performance of the system, namely, frame-based accuracy and frame-based F score. The frame-based accuracy is defined as

$$accuracy = \frac{number\ of\ correctly\ labeled\ frames}{total\ number\ of\ frames} \quad . \quad (4)$$

The frame-based F score is defined as the harmonic mean of precision and recall values. Precision is the ratio of true positive value to the test outcome positives for a particular class. True positive value is the number of frames correctly labeled by the system for a particular class, and test outcome positive value is the number of frames detected by the system belonging to that class. Recall is the ratio of true positive values to total positive values for any class. Total positive values are number of frames in the test set belonging to that particular class. The frame-based F score is thus given by

$$F\text{score} = 2\,\frac{P \cdot R}{P + R} \quad , \quad (5)$$

where P and R are the precision and recall, respectively. Accuracy provides the overall performance of the system, while F score is a measure of performance over individual classes. The proposed segmentation system aims to identify expiratory and inspiratory phases from cry recordings; hence, F scores are reported for these two classes only. The final system performance metrics were obtained by averaging results over all five folds.

It is to be noted that each performance metric is accompanied with the standard error calculated as,

$$\text{standard error} = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}} \tag{6}$$

where the sample sizes for Tampere and Cape Town cohorts are 57 and 52, respectively.

5.1 Results: Tampere cohort

We investigated the performance of the system with changes in the following parameters of the system,

- Number of HMM states used for each target class.
- Number of Gaussian components used to model the output of each HMM state.
- Audio features used for feature extraction.

We started with a baseline HMM configuration consisting of one state and five Gaussian components for each target class while using standard MFCCs. Figure 10 (left plot) shows the variation of system accuracy and F scores while increasing the number of HMM states. The number of Gaussian components for each target class is 5. It can be seen that increasing the number of HMM states up to 3 leads to improvement in the system performance. However, adding further states does not result in any significant improvement in system performance but results in an increase in computation time of training the HMMs.

Similarly, Fig. 10 (right plot) depicts the variation in system performance while increasing the number of Gaussian components. HMMs with one state for all target classes were used. It shows an improvement in system accuracy on incorporating up to 15 Gaussian components. On incorporating more number of Gaussians, improvement is not very substantial; on the other hand, computation time of training HMM also increases.

Tables 2 and 3 show the accuracy and F scores, respectively, for different combination of HMM states and component Gaussians. As we add further states and Gaussian components into the HMM topology, we improve the system performance, but at the same time, training time of the models increases as well. The solution would be to choose a topology which is fairly efficient in terms of segmentation metrics and does not take much time to train. For further experiments, we empirically chose a topology consisting of three HMM states for each class and 10 Gaussian components to model each HMM state.

It can be observed that performance of the segmentation system is good for expiratory phases, while it is relatively poor for inspiratory phases. This observation can be attributed to short duration, lack of training data, and wide variation in the types of inspiratory phases present in the database, as was pointed out in Section 4.1. The number of instances of expiratory phases in the database, including both Tampere and Cape Town data, is around 1.7 times larger than inspiratory phases.

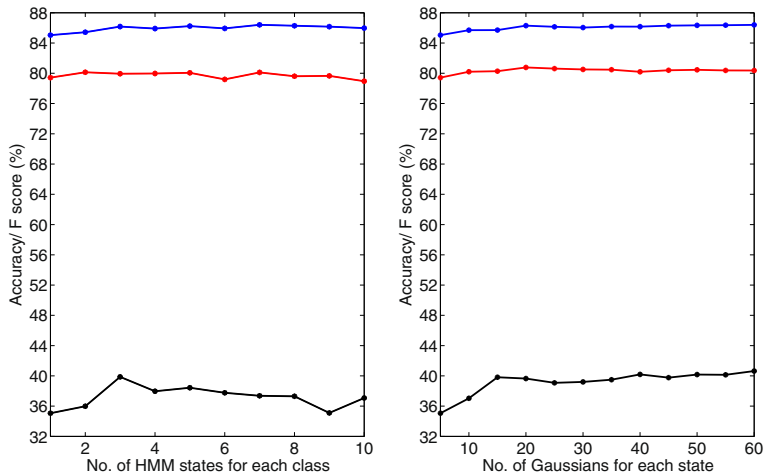

Fig. 10 The left plot shows the variation of segmentation performance for different number of HMM states per target class, using five Gaussians for each state. The right plot shows the same for different number of Gaussians for each state, using single HMM state per target class. The blue plot depicts the variation in system accuracy, while red and black plots depicting variation in F scores for expiratory and inspiratory phases, respectively

Table 2 Accuracy (in %) of the audio segmentation system with different number of HMM states and component Gaussians

No. of Gaussians	No. of HMM states for each class			
	1	2	3	4
5	85.0 ±1.3	87.1 ±1.2	87.8 ±1.2	87.7 ±1.1
10	85.7 ±1.2	85.4 ± 1.1	87.5 ±1.1	87.6 ± 1.1
15	85.7 ± 1.2	85.5 ± 1.0	87.9 ±1.1	87.9 ±1.0
20	86.1 ±1.1	87.8 ± 1.0	87.8 ±1.0	88.1 ±1.1
25	86.1 ±1.1	88.0 ±1.0	87.9 ±1.0	88.2 ±1.1

Table 4 reports the performance of the system with additional features. An improvement in the system performance is observed by combining other audio features with MFCCs used in the baseline configuration. Use of deltas and delta-deltas, F0, and aperiodicity features led to an overall improvement in the accuracy of the system. A corresponding improvement in the *F* score performance was observed as well, notably for inspiratory phases. The overall accuracy of the system was improved up to 89.2% for a combination of MFCCs and aperiodicity features. The corresponding *F* score performance was 48.9% for inspiratory phases and 83.3% for expiratory phases.

5.2 Cape Town cohort

The recording conditions (e.g., acoustic environment and recording equipment) while capturing cry data in Tampere and Cape Town were quite different from each other. In this section, we report the ability of the system trained on Tampere data to work for Cape Town data using techniques discussed in Section 3. The effectiveness of employed adaptation techniques was investigated by comparing the performance of the adapted system with the system trained on Tampere data alone and with the one trained on Cape Town data (without any adaptation).

The segmentation system was trained on Tampere data (entire 57 recordings) and tested on Cape Town data (entire 52 recordings). The output labels generated by the system were compared against manually annotated ground truth obtained for Cape Town cohort. The final system performance metrics were obtained by averaging over all Cape Town recordings. An accuracy of 58.3% was obtained for this system. For model adaptation, this procedure was repeated, firstly, with feature normalization alone, and then, with both feature normalization and semi-supervised learning adaptation. Table 5 compares the performance of the original system (trained on Tampere data and tested on Cape Town data) with that of the adapted system. Most of the improvement was achieved through feature normalization step. It can be seen that feature normalization improved the system accuracy to 80% from 58.3% for the original system (trained on Tampere material alone). Semi-supervised learning further improved it with a notable improvement in *F* score for expiratory phases up to 73.2% against 71.2% with feature normalization alone. Similar improvements in *F* score of inspiratory phases was observed, as reported in Table 5.

Instead of using the entire unlabeled data for semi-supervised learning adaptation, a confidence threshold can be used to select only a subset of the unlabeled data, as explained in Section 3.2.

6 Conclusions

In this paper, we investigate the problem of automatically identifying expiratory and inspiratory phases from infant cry recordings. The segmentation system offers system accuracies up to 89.2% and is capable of adapting to cry sounds recorded in acoustic settings different from the one it is trained for.

Two datasets, Tampere cohort with 57 cry recordings and Cape Town cohort with 52 recordings, were analyzed. The recordings were captured under realistic clinical environments which often consisted of extraneous sound sources. The output of this segmentation system can then be utilized for performing further analysis involving extraction of required acoustic parameters from the identified acoustic parts. This segmentation system thus offers to be an essential pre-processing step for an infant cry analysis system especially when the number of cry recording to be analyzed is large enough to render manual segmentation unfeasible.

Table 3 *F* scores (%) of the expiratory and inspiratory phase classes, denoted by Exp and Ins, respectively, with different number of HMM states and component Gaussians

No. of Gaussians for each state	No. of HMM states for each class							
	1		2		3		4	
	Ins	Exp	Ins	Exp	Ins	Exp	Ins	Exp
5	35.1 ±2.6	79.4 ±1.6	38.3 ±2.6	81.1 ±1.6	41.0 ±2.7	81.1 ±1.6	40.5 ±2.7	80.8 ±1.5
10	37.0 ±2.6	80.2 ±1.6	40.1 ±2.4	81.4 ±1.4	41.5 ±2.8	80.6 ±1.4	40.8 ±2.7	81.4 ±1.5
15	39.8 ±2.6	80.3 ±1.5	41.6 ±2.5	81.7 ±1.2	42.8 ±2.6	80.5 ±1.3	41.8 ±2.9	80.7 ±1.3
20	39.6 ±2.6	80.7 ±1.5	41.7 ±2.7	81.6 ±1.3	40.6 ±2.8	81.2 ±1.3	41.7 ±2.9	81.2 ±1.4
25	39.8 ±2.6	80.6 ±1.4	42.3 ±2.6	81.9 ±1.3	41.3 ±2.8	81.0 ±1.3	42.4 ±2.8	80.9 ± 1.3

Table 4 The performance of the model with additional features

Features	Acccuracy	*F* score (%)	
	(%)	Ins	Exp
MFCCs	87.5 ±1.1	41.5 ±2.8	80.6 ±1.4
MFCCs + Δ and ΔΔ	88.6 ±0.9	45.9 ±2.5	82.2 ±1.1
MFCCs + *F*0	88.5 ±0.9	47.8 ±2.9	81.6 ±1.3
MFCCs + *F*0 + Δ and ΔΔ	88.8 ±0.9	47.1 ±2.5	82.3 ±1.1
MFCCs + ap0	89.2 ±0.9	48.9 ±2.8	83.3 ±1.3
MFCCs + ap0+ Δ and ΔΔ	89.0 ± 0.8	47.4 ±2.5	82.8 ±1.2

The expiratory and inspiratory phase classes are denoted by Exp and Ins, respectively

The cry recordings utilized in this study were captured under a wide variation in the recording conditions (i.e., context of recording, type of cry trigger, and types of extraneous sound sources present while recording). Moreover, infant-related attributes known to affect acoustic characteristics of cry (e.g., weight of the infant, prematurity of birth) varied as well.

An HMM based solution is proposed for the segmentation problem. The cry recordings were segmented into three classes: expiratory phases, inspiratory phases, and residual. The former two classes constitute the regions of interest, and residual is simply a collection of all irrelevant acoustic regions (i.e., non-cry vocals of infant, other sound sources, and silent parts). The HMM configuration, namely, the number of states for each class and the number of Gaussian components used to model each HMM state, were varied, and the resulting effect on system performance was investigated. An improvement in system performance was observed while using more than one HMM state for each class and adding more component Gaussians. However, the improvement was not very significant beyond certain number of states and component Gaussians, and as we opted for a larger number of HMM states and number of Gaussian components, the computation times for training also increased. Hence, in the proposed system, a suitable number of HMM states and component Gaussians should be chosen based on availability of training material and requirements of computation time. In this study, we have presented results for three HMM states for each class and 10 component Gaussians for modeling each HMM state.

Table 5 Comparison of the performance of the original segmentation system with the adapted system on Cape Town database

	Accuracy	*F* score (%)	
	(%)	Insp	Exp
Original system	58.3 ±1.4	20.7 ±0.7	65.4 ± 1.6
With feature Norm. alone	80.0 ±1.4	36.5 ±2.0	71.2 ±1.6
With feature norm. and SSL	80.7 ±1.4	38.5 ± 1.8	73.2 ±1.7
System trained with Cape Town data	85.2 ±0.7	39.2 ±2.3	78.0 ± 1.1

The expiratory and inspiratory phase classes are denoted by Exp and Ins, respectively

It is observed that the segmentation system works sufficiently well for expiratory phases but performs rather poorly for inspiratory phases in comparison. The reason for this is the diverse nature of inspiratory phases present in our data set. Additionally, less data were available for training the system to identify inspiratory phases as compared to expiratory phases. The performance of the system for inspiratory phases is expected to improve with the availability of more training material.

Different audio features in conjunction with conventional MFCC features were experimented with. An improvement in system performance was observed with deltas and delta-deltas, fundamental frequency, and aperiodicity features. It is hence recommended to incorporate them along with MFCCs. The best performance for the system was observed with aperiodicity feature. An accuracy of 89.2% along with *F* scores of 48.9 and 83.3% were obtained for inspiratory and expiratory phases, respectively.

As a critical test of the applicability of the proposed segmentation system to cry recordings irrespective of recording conditions, we show that the system trained on material recorded in one acoustic setting can be reliably adapted to perform on material recorded in an unseen acoustic setting. We propose a two-step model adaptation method consisting of feature normalization and semi-supervised learning adaptation. The proposed model adaptation method yielded system accuracies up to 80.7% compared to 85.2% obtained for a system trained on the material recorded in the unseen acoustic setting itself.

In this study, we have grouped together all extraneous acoustic parts in a single target class called residual. Alternatively, multiple classes can be created for different kinds of sound sources provided sufficient data is available. Moreover, in this study, HMM topology with same number of states for all target classes were used. Alternatively, different combinations of number of states for different classes can be experimented with for optimal performance. Additionally, instead of using the self-training method, other semi-supervised learning methods can be experimented with in order to improve the ability of the system to adapt to material captured in unseen acoustic settings.

Acknowledgements
The study presented here was supported by a joint research grant from the Academy of Finland and National Research Foundation, South Africa. We express our gratitude to the staff of Neonatal Ward Unit, Neonatal Intensive Care Unit, and Rooming-in Ward Unit of Tampere University Hospital, and Intercare Well-Baby Clinic in Cape Town for their cooperation in collection of cry recordings used in this study. We also gratefully acknowledge the help of Dana Niehaus and Gerdia Harvey in collecting the data for the Cape Town cohort.

Authors' contributions
GN and JK share joint first authorship of this article. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 10, Tampere, Finland. [2]School of Medicine, University of Tampere, Kalevantie 4, Tampere, Finland. [3]Department of Pediatrics, Tampere University Hospital, Tampere, Finland. [4]School of Social Sciences and Humanities, University of Tampere, Tampere, Finland.

Received: 11 May 2017 Accepted: 15 January 2018
Published online: 26 January 2018

References

1. MJ Corwin, BM Lester, C Sepkoski, M Peucker, H Kayne, HL Golub, Newborn acoustic cry characteristics of infants subsequently dying of sudden infant death syndrome. Pediatrics. **96**(1), 73–77 (1995)
2. HL Golub, MJ Corwin, Infant cry: a clue to diagnosis. Pediatrics. **69**(2), 197–201 (1982)
3. C Manfredi, M D'Aniello, P Bruscaglioni, A Ismaelli, A comparative analysis of fundamental frequency estimation methods with application to pathological voices. Med. Eng. Phys. **22**(2), 135–147 (2000)
4. K Michelsson, O Michelsson, Phonation in the newborn, infant cry. Int. J. Pediatr. Otorhinolaryngol. **49**, 297–301 (1999)
5. G Esposito, P Venuti, M Bornstein, Assessment of distress in young children: a comparison of autistic disorder, developmental delay, and typical development. Res. Autism Spectrum Disorders. **5**(4), 1510–1516 (2011)
6. LL LaGasse, AR Neal, BM Lester, Assessment of infant cry: acoustic cry analysis and parental perception. Ment. Retard. Dev. Disabil. Res. Rev. **11**(1), 83–93 (2005)
7. L Rautava, A Lempinen, S Ojala, R Parkkola, H Rikalainen, H Lapinleimu, L Haataja, L Lehtonen, PS Group, *et al*, Acoustic quality of cry in very-low-birth-weight infants at the age of 1 1/2 years. Early Hum. Dev. **83**(1), 5–12 (2007)
8. P Zeskind, M McMurray, CL ET, K Grewen, K Garber, J Johns, Translational analysis of effects of prenatal cocaine exposure on human infant cries and rat pup ultrasonic vocalizations. PloS ONE. **9**(10), 110349–110349 (2013)
9. A Fort, C Manfredi, Acoustic analysis of newborn infant cry signals. Med. Eng. Phys. **20**(6), 432–442 (1998)
10. BM Lester, Developmental outcome prediction from acoustic cry analysis in term and preterm infants. Pediatrics. **80**(4), 529–534 (1987)
11. S Orlandi, C Manfredi, L Bocchi, M Scattoni, in *Proc. Annual Intl. Conf. IEEE Engineering in Medicine and Biology Society*. Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis (IEEE, 2012), pp. 2953–2956
12. AM Chilosi, P Cipriani, C Pecini, D Brizzolara, L Biagi, D Montanaro, M Tosetti, G Cioni, Acquired focal brain lesions in childhood: effects on development and reorganization of language. Brain Lang. **106**(3), 211–225 (2008)
13. F Cunha, JJ Heckman, The economics and psychology of inequality and human development. J. Eur. Econ. Assoc. **7**(2-3), 320–364 (2009)
14. PS Douglas, PS Hill, A neurobiological model for cry-fuss problems in the first three to four months of life. Med. Hypotheses. **81**(5), 816–822 (2013)
15. M Hadders-Algra, Challenges and limitations in early intervention. Dev. Med. Child Neurol. **53**(s4), 52–55 (2011)
16. KE Pape, Developmental and maladaptive plasticity in neonatal SCI. Clin. Neurol. Neurosurg. **114**(5), 475–482 (2012)
17. AM Goberman, MP Robb, Acoustic examination of preterm and full-term infant cries: The long-time average spectrum. J. Speech Lang. Hearing Res. **42**(4), 850–861 (1999)
18. K Wermke, MP Robb, Fundamental frequency of neonatal crying: does body size matter? J. Voice. **24**(4), 388–394 (2010)
19. A Branco, SM Fekete, LM Rugolo, MI Rehder, The newborn pain cry: descriptive acoustic spectrographic analysis. Int. J. Pediatr. Otorhinolaryngol. **71**(4), 539–546 (2007)
20. FL Porter, SW Porges, RE Marshall, Newborn pain cries and vagal tone: parallel changes in response to circumcision. Child Dev. **59**, 495–505 (1988)
21. A Messaoud, C Tadj, in *Proc. Intl. Conf. on Image and Signal Processing*. A cry-based babies identification system (Springer, 2010), pp. 192–199
22. B Reggiannini, SJ Sheinkopf, HF Silverman, X Li, BM Lester, A flexible analysis tool for the quantitative acoustic assessment of infant cry. J. Speech Lang. Hearing Res. **56**(5), 1416–1428 (2013)
23. G Várallyay, A Illényi, Benyó, in *Proc. Intl. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*. Automatic infant cry detection, (2009), pp. 11–14
24. S Orlandi, L Bocchi, G Donzelli, C Manfredi, Central blood oxygen saturation vs crying in preterm newborns. Biomed. Signal Process. Control. **7**(1), 88–92 (2012)
25. S Orlandi, PH Dejonckere, J Schoentgen, J Lebacq, N Rruqja, C Manfredi, Effective pre-processing of long term noisy audio recordings: an aid to clinical monitoring. Biomed. Signal Process. Control. **8**(6), 799–810 (2013)
26. R Cohen, Y Lavner, in *Proc. IEEE 27th Convention of Electrical and Electronics Eng. in Israel*. Infant cry analysis and detection (IEEE, 2012), pp. 1–5
27. SM Grau, MP Robb, AT Cacace, Acoustic correlates of inspiratory phonation during infant cry. J. Speech Lang. Hear. Res. **38**(2), 373–381 (1995)
28. J-J Aucouturier, Y Nonaka, K Katahira, K Okanoya, Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models. J. Acoust. Soc. Am. **130**(5), 2969–2977 (2011)
29. L Abou-Abbas, HF Alaie, C Tadj, Automatic detection of the expiratory and inspiratory phases in newborn cry signals. Biomed. Signal Process. Control. **19**, 35–43 (2015)
30. L Abou-Abbas, C Tadj, C Gargour, L Montazeri, Expiratory and inspiratory cries detection using different signals' decomposition techniques. J. Voice. **31**(2), 259–13 (2017)
31. D Lederman, E Zmora, S Hauschildt, A Stellzig-Eisenhauer, K Wermke, Classification of cries of infants with cleft-palate using parallel hidden Markov models. Med. Biol. Eng. Comput. **46**(10), 965–975 (2008)
32. D Lederman, A Cohen, E Zmora, K Wermke, S Hauschildt, A Stellzig-Eisenhauer, in *Proc. IEEE 22nd Convention of Electrical and Electronics Eng. in Israel*. On the use of hidden markov models in infants' cry (IEEE, 2002), pp. 350–352
33. D Reynolds, in *Encyclopedia of biometrics*. Gaussian mixture models (Springer, USA, 2009), pp. 659–663
34. L Rabiner, B-H Juang, An introduction to hidden Markov models. ASSP Mag. IEEE. **3**(1), 4–16 (1986)
35. M Gales, S Young, The application of hidden Markov models in speech recognition. Foundations Trends Signal Process. **1**(3), 195–304 (2008)
36. S Davis, P Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoust. Speech Signal Process. IEEE Trans. **28**(4), 357–366 (1980)
37. JT Foote, in *Voice, Video, and Data Communications*. Content-based retrieval of music and audio (International Society for Optics and Photonics, 1997), pp. 138–147
38. O-W Kwon, K Chan, J Hao, T-W Lee, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*. Emotion recognition by speech signals (Citeseer, 2003), pp. 125–128
39. De Cheveigné, H Kawahara, YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002)
40. De Cheveigné, H Kawahara, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*. Comparative evaluation of F0 estimation algorithms, (2001), pp. 2451–2454
41. A von dem Knesebeck, U Zölzer, in *Proc. Int. Conf. Digital Audio Effects*. Comparison of pitch trackers for real-time guitar effects, (2010), pp. 266–269

42. YIN pitch estimator. http://audition.ens.fr/adc/sw/yin.zip. Accessed 2 Feb 2015
43. L Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. **77**(2), 257–286 (1989)
44. SJ Young, S Young, The HTK hidden Markov model toolkit: design and philosophy. Entropic Cambridge Res. Lab. Ltd. **2**, 2–44 (1994)
45. Daumé, A Course in Machine Learning (2014). http://ciml.info/. Accessed 2 Feb 2015
46. X Zhu, Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin-Madison (2005)
47. Audacity Team, Audacity(R): free audio editor and recorder [Computer program] (2014). Version 2.0.3. http://audacity.sourceforge.net/. Accessed 2 Feb 2015