Biomedical Signal Processing and Control 19 (2015) 35–43

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Automatic detection of the expiratory and inspiratory phases in newborn cry signals

Lina Abou-Abbas*, Hesam Fersaie Alaie, Chakib Tadj

Electrical Engineering Department, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

ARTICLE INFO

Article history:
Received 31 July 2014
Received in revised form 12 March 2015
Accepted 13 March 2015
Available online 3 April 2015

Keywords:
HMM
Automatic segmentation
Newborn cry signals
Mel Frequency Cepstral Coefficients
Viterbi algorithm
Baum Welch algorithm

ABSTRACT

An analysis of newborn cry signals, either for the early diagnosis of neonatal health problems or to determine the category of a cry (e.g., pain, discomfort, birth cry, and fear), requires a primary and preliminary preprocessing step to quantify the important expiratory and inspiratory parts of the audio recordings of newborn cries. Data typically contain clean cries interspersed with sections of other sounds (generally, the sounds of speech, noise, or medical equipment) or silence. The purpose of signal segmentation is to differentiate the important acoustic parts of the cry recordings from the unimportant acoustic activities that compose the audio signals. This paper reports on our research to establish an automatic segmentation system for newborn cry recordings based on Hidden Markov Models using the HTK (Hidden Markov Model Toolkit). The system presented in this report is able to detect the two basic constituents of a cry, which are the audible expiratory and inspiratory parts, using a two-stage recognition architecture. The system is trained and tested on a real database collected from normal and pathological newborns. The experimental results indicate that the system yields accuracies of up to 83.79%.

1. Introduction

With early newborn screening, a serious illness can be diagnosed such that treatment can begin before severe problems appear, and in certain cases, sudden mortality or disability can be prevented. Clearly, the presence of disease must be detected at an early stage. Systematic screening combined with better diagnostic tools is therefore required to meet future medical challenges, with the aim of supporting clinical decision-making and improving the effectiveness of treatment [1]. These tools have evolved considerably in recent years in terms of improving screening and symptom evaluation, and the newborn cry signal has been the object of strong research interest for the past three decades.

Researchers have amassed enough evidence to conclude that a cry signal contains relevant information on the psychological and physiological condition of the newborn, formal relationships have been established between the acoustic features extracted from the cries and the health problems of the child [2–5]. Various studies are currently under way to devise a tool that analyzes cries automatically, to diagnose neonatal pathologies [6–8].

We are involved in the design of an automatic system for early diagnosis, called the Newborn Cry-based Diagnostic System (NCDS), which can detect certain pathologies in newborns at an early stage. The implementation of this system requires a database containing hundreds of cry signals.

The overwhelming problem that arises when working with such a database is the diversity of acoustic activities that compose the audio recordings, such as background noise, speech, the sound of medical equipment and silence. Such diversity could harm the analysis process, as the presence of any acoustic component other than the cry itself could result in the misclassification of pathologies by reducing the NCDS system performance. This is because the NCDS would decode every segment of the recording signal, whether it is part of a cry or not. In this case, unwanted segment insertion in essential crying segments would lengthen the process of classification unnecessarily and leave the system prone to error. An important subtask of the NCDS is the manipulation of the newborn cry sound, and what is needed to perform this subtask is a segmentation system. Until now, few works have been carried out in this area. In this paper, we propose an automatic segmentation module designed to isolate the audible expiration and inspiration parts of cry sounds to serve as a preprocessing step of our NCDS.

The rest of this paper is organized as follows: Related work is presented in Section 2. The HMM and the HTK are reviewed briefly in Section 3. The training corpus and the testing corpus are described in Section 4. In Section 5, the architecture of the

* Corresponding author. Tel.: +1 5144736060.
E-mail addresses: Lina.Abou-Abbas.1@etsmtl.net (L. Abou-Abbas), Hesam.farsaie-alaie@etsmtl.net (H. Fersaie Alaie), Chakib.Taj@etsmtl.ca (C. Tadj).

http://dx.doi.org/10.1016/j.bspc.2015.03.007
1746-8094/

proposed system is presented, and details of the individual blocks are described in five subsections. Section 6 contains the implementation of the system, the obtained results, and the discussion. Finally our conclusions are presented in Section 7.

2. Related work

Several studies have been conducted in which the infant cry is analyzed (categorization of the cry, disease classification based on the cry). In 1985, for example, Corwin and Golub outlined four acoustic categories composing a cry episode, which are: (a) expiratory phonation (with F0 ranging from 250 to 750 Hz), (b) expiratory hyperphonation (with F0 ranging from 1000 to 2000 Hz), (c) expiratory dysphonation (aperiodic expiratory segment), (d) inspiratory phonation (associated with any perceptually audible sound generated by the newborn during inspiration, or high-pitched cries during inspiration) [2,9].

In most studies, the cry segmentation phase was performed manually, a human operator was asked to monitor the recorded audio signals and pick out only the important cry parts from the recordings [3,10,11]. This manual task is tiresome and too time-consuming when the volume of data is large. The cry segmentation that serves the needs of a real-time diagnostic tool should be performed automatically.

In some studies, the authors have applied various voice activity detection software approaches such as the traditional methods of ZCR (Zero Crossing Rate) and STE (Short Time Energy), with some modification of the thresholds [4,5,12,13]. In general, these methods are of limited use in this context, as speech and cry sounds have different features. With these methods, particularly in the search for the high-energy parts of the audio signals, not only are the meaningful parts of cry vocalizations found but also background noise, speech, and machine sounds. In other words, the typical voice activity detection methods alone are not suitable for segmenting a cry signal. The corpuses used to examine these methods (ZCR, STE) were composed only of cry sounds, which are sequences of expiration and inspiration, alternating with short periods of silence and background noise. The main goal of authors was to eliminate silence and background noise without affecting the audible expiration and inspiration phases.

Few studies have been conducted specifically on the automatic segmentation of cry signals [14–17]. Two novel algorithms were introduced by modifying the Harmonic Product Spectrum (HPS) method [14]. The HPS method was created to detect the fundamental frequency of an audio signal. The authors showed that it is possible to check the regularity structure of the spectrum using the HPS method and classify its content by detecting the meaningful parts of the cry sounds. Another study on the segmentation of cry signals was conducted by Cohen [16] with the purpose of labeling each successive segment as a cry/non-cry/non-activity. However, with the methods presented in [16], the inspiration parts as well as the dysphonic vocalizations of the cry spectrum that could be presented with irregular or non-harmonic structure were ignored.

Recent studies have shown that differentiated characteristics in expiratory and inspiratory vocalizations exist in adults as well in newborns [18].

Assuming that the inspiratory phase of a cry episode reflects a laryngeal contraction of the ingressive airstream, inspiratory vocalization has been proven to be useful in the identification of newborns at risk for various health conditions [9]. In fact, the amount of time the inspiratory phase lasts in newborns with respiratory disorders is greater than it is in normal newborns [19]. Indeed, recent medical evidence confirms that a relationship exists between upper airway obstruction and sudden infant death syndrome and sleep apnea. Despite this evidence, it is surprising to find acoustic data that are limited to the expiratory phase alone [9].

To create an effective diagnostic tool based on the cry signals, the involvement of both the expiratory and the inspiratory components is a prerequisite. The aim of this study is to identify and quantify both the audible inspiratory and expiratory components of a newborn cry automatically.

The work presented in this paper is based on the well-established and widely used Hidden Markov Model (HMM) statistical technique, which has been successfully applied in automatic speech recognition and segmentation systems.

To the best of our knowledge, no work has yet been carried out on the automatic segmentation of crying signals recorded in noisy environments without manually pre-processing the signals to remove at least irrelevant acoustic activities, such as speech and beep sounds around the infant.

In recent work [17], authors applied an automatic segmentation approach based on a HMM classification tool to segment the expiratory and inspiratory sounds of cry signals. The difference with this recent approach compared to our approach is not only with the limited number of infants and the limited available acoustic activities types (due to the environment in which recordings are taking place) but also the way in which they applied the HMM. The authors of [17] considered only three classes, Expiration (EX), Inspiration (IN) and Silence (SI). As a first stage, to train each class, they used different techniques such as Support Vector Machines (SVM) as well as Gaussian Mixture Models (GMM) consisting of 5 and 20 Gaussian components. To reduce errors by taking into account the arrangement in time between the three classes, the authors added a second stage using the Viterbi algorithm. The whole architecture of the approach in [17] could be taken as an HMM architecture of three states. In fact, the segmentation approach presented in [17] performed well, but its performance needs to be enhanced to segment audio signals recorded in a noisy environment (e.g., sounds of speech, medical equipment, noise, and silence).

To provide a better understanding of the context of this study, some important terms used must be predefined.

- *Inspiration* is associated with inspiratory phonation as defined by Golub and Corwin [2].
- *Expiration* is referred to the acoustic output during the expiration phase of a cry (it can be phonation, dysphonation, or hyperphonation), as well as any audible expiration sound generated by the infant outside its cry episodes. Note that we do not make a distinction here between the expiration phases that occur during or following a cry.
- *A cry sequence* consists of long periods of expiratory crying separated by short inspiratory episodes.

We have avoided using the terms voiced inspiration and voiced expiration to describe the important parts of the cry. In fact, a dysphonation vocalization is characterized in earlier studies as an unvoiced part during a cry and it is considered one of the most useful vocalizations in the detection of newborns at risk of various health conditions [20]. For this reason, we prefer using the terms audible inspiration and audible expiration.

3. Hidden Markov Model and the HTK

HMMs underlie the most modern automatic speech recognition (ASR) systems. They have many potential applications in statistical signal processing and acoustic modeling, including the segmentation of recorded signals [21]. The basic principles of any ASR system involve constructing and manipulating a series of statistical models that represent the various acoustic activities of the sounds to

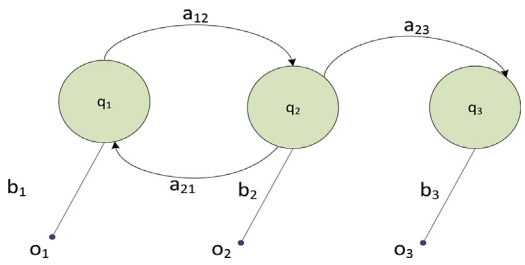


Fig. 1. Hidden Markov Model topology.

be recognized [21]. Many studies have shown that speech, music, newborn cries, and other sounds can be represented as a sequence of feature vectors (temporal, spectral, or both), and HMMs could provide a very important and effective framework for building and implementing time-varying spectral vector sequences [22].

An HMM generates a sequence of observations $O = O_1, O_2, \ldots, O_T$ and is defined by the following parameters: number of hidden states, state transition probability distribution A, observation probability distribution B, and initial state distribution π. We denote the model parameters of the HMM as $\lambda = \{A, B, \pi\}$ [4,12]. These concepts are depicted in Fig. 1.

To build and manipulate an HMM, three problems must be solved: the evaluation problem, the decoding problem, and the training problem. HMM theory, the aforementioned problems, and the proposed solutions are widely explained in the literature, especially in the well-known Rabiner tutorial [23]. The Viterbi algorithm is proposed as a decoding solution to find the most probable future state of the system based on its current state [23]. The Baum Welch algorithm is an iterative procedure used to estimate the HMM parameters.

The HTK is the Hidden Markov Model Toolkit developed by Steve Young in the Cambridge University Engineering Department (CEUD). This toolkit is designed to build and manipulate HMMs using training observations from a sound corpus to decode unknown observations. It consists of a set of library modules and tools available in the C source code. Although the use of the HTK has been limited to speech recognition research, it is flexible enough to support the development of various HMM systems [21].

4. Materials

To develop our targeted diagnostic system (NCDS), we are working on building a very large corpus of cry signals. The recordings were made in the neonatology departments of several hospitals in Canada and Lebanon. The infants that were selected for the recording procedures are from 1 to 53 days old and were both preterm and full term. The database includes both healthy and sick babies and both males and females. The average duration of the newborn cry records is 90 s. The medical staff was put in charge of the following tasks: determining the type of cry being recorded, such as pain, hunger, diaper change and birth cry, writing down the date and time of the cry recording, and any useful information available about the babies (date of birth, gender, maturity, race, ethnicity, gestation, and known diseases). Three recording files were collected from the majority of the babies. All the recordings were acquired with an Olympus hand-held digital 2-channel recorder at a sampling frequency of 44.1 kHz and a sample resolution of 16 bits. The recorder was placed 10 to 30 cm from the newborn's mouth to be effective. The recorded audio signals are registered as WAV files. The newborns that were selected for the global NCDS project suffer from various pathological conditions. The group of abnormal infants represents various types of serious conditions and diseases, chief among them being:

Table 1
Corpus statistics.

			Number of babies	Number of signals
Male	Full term	Healthy	3	8
		Pathological	3	7
	Preterm	Healthy	3	6
		Pathological	4	7
Female	Full term	Healthy	22	50
		Pathological	12	31
	Preterm	Healthy	7	16
		Pathological	10	26
Total			64	151

- Diseases affecting the central nervous system (cerebral hemorrhage, meningitis, sepsis).
- Blood disorders (anemia, hyperbilirubinemia, hemolytic disease, hypoglycemia).
- Congenital cardiac anomaly (ventricular septal defect, atrial septal defect, complex cardiovascular cases).
- Diseases in which the respiratory system is directly involved (asphyxia, respiratory distress syndrome, apnea, bronchopulmonary dysplasia, and pneumonitis).
- Chromosomal abnormality.

Thus far the corpus collected includes infants' cries in different recording environments and conditions, from silent to very noisy. The background noises may be of many types, such as human speech (nurses, doctors, parents), the sounds of the recording device and medical equipment in the neonatal Intensive Care Unit (the beeping of machines), and the sounds made of doors opening and closing and running water. To build our segmentation module, we used signals produced by 64 newborns, including both normal and abnormal, for a total of 151 cry signals (Table 1).

The total duration of the recordings in the training corpus and the testing corpus used here is 21,900 s: 4 h and 11 min are devoted to the training samples, and 1 h and 54 min for the testing samples. It is important here to note that the babies chosen for the training phase are different from those chosen for the testing phase.

This content of this database is unique and realistic. It contains long cry sequences (expiration phases alternating with short periods of inspiration episodes). The cry signals in both the training corpus and the testing corpus have been manually indexed – see Table 2 for details.

5. System overview

In this paper, we introduce a new approach for high performance cry signal segmentation on realistic tasks related to our automatic Newborn Cry-based Diagnostic System (NCDS). We use a Cygwin interface, which enables HTK commands to be executed on the Windows platform. This segmentation tool is designed mainly to automatically segment and label expiration and inspiration phases

Table 2
Data used for training and testing corpuses.

	Training corpus in min	Testing corpus in min	Total time in min
EX	72.1	38.7	110.8
IN	8.1	3.6	11.7
SP	8.6	6.3	14.9
BIP	2.8	1.8	4.6
SI	58.4	29.5	87.9
NS	27.7	14.7	42.4

EX – Expiration; IN – Inspiration; SP – Speech; BIP – Beeping of machines; SI – Silence; NS – Noise.

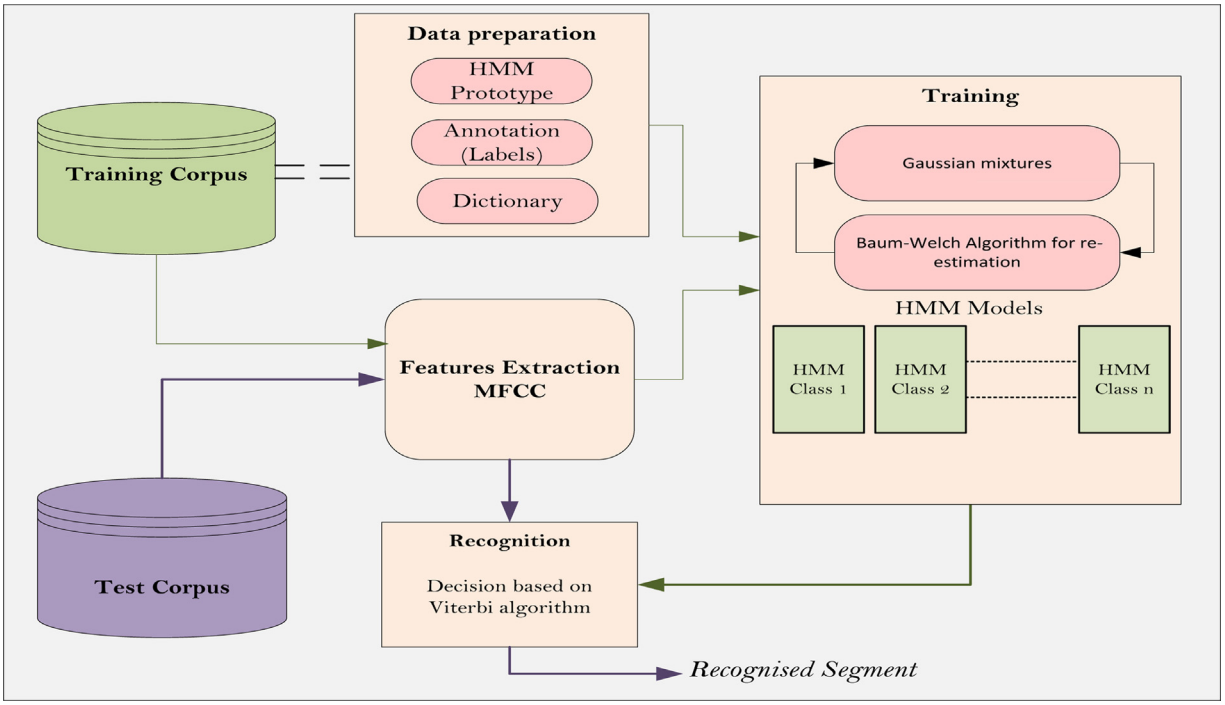


Fig. 2. Automatic infant cry segmentation system architecture.

taken from the audio recordings of newborn cries, in an attempt to find the best ways to resolve these two major issues:

a) Select the best parameters of the MFCC extraction procedure such as the number of MFCCs and the window size.
b) Design a robust classifier that can perform accurate segmentation. Cry signal segmentation using the HMM approach can be viewed as similar in its implementation to cry signal classification problems (classification of diseases or types of cry).

To implement the HMM efficiently, we used a Cygwin interface and the HTK. The cry segmentation system architecture consists of the following modules: data preparation, feature extraction, training, and recognition. A block diagram of the proposed system is shown in Fig. 2. Individual blocks are described in the following subsections.

5.1. Data preparation

The first step in any recognizer tool is data preparation, as data are needed for both training and testing. This phase consists of recording and labeling the cry signals. The recording task was discussed in Section 4. Labeling is required so that the HMM models can be trained and the results of the proposed automatic segmentation system can be tested. For this task, which is performed manually, we used the Wave Surfer software [24]. For the labeling task, a task dictionary in text format is required to describe the correspondence between the name of the class and the label. These text files are saved in the .lab format. An example of manual segmentation is shown in Fig. 3. We chose to build 6 HMMs, according to the various sound activities recorded, as follows:

- *Expiration class (EX)*: composed of the acoustic activities of the baby during expiration episodes.
- *Inspiration class (IN)*: composed of the vocal activity of the baby during the inspiration phase.
- *Noise class (NS)*: composed of the sounds produced by the recording device and background sounds.
- *Speech class (SP)*: composed of the sounds made by speakers within the recording area.
- *Silence class (SI)*: composed of periods of lack of sound.
- *Bip Class (BIP)*: the sounds made by medical equipment, characterized by uniform energy.

5.2. Features extraction

The cry signals captured by the recorder are fed into the Feature Extraction module. At this stage, the input signal is first converted into a series of acoustic vectors, which are then be forwarded to the Recognition (decoding) phase [25]. In the feature extraction phase, the audio signals are expressed in spectral form by converting the raw audio signal into a sequence of acoustic feature vectors that carry acoustic information about the signal. MFCC (Mel Frequency Cepstral Coefficients) is one of the most efficient and widely used parameterization techniques used to produce the feature vectors. The audio waveforms, sampled at 44.1 kHz, are treated with a pre-emphasis coefficient of 0.97. The Hamming window is then applied, and then the Fourier Transform is calculated for each frame.

The obtained power spectrum is fed to the Mel-scale filter banks (24 channels) to yield more low frequency details, and the Mel coefficients are generated by applying the Discrete Cosine Transformation (DCT) to the log spectral representation (see Fig. 4). MFCC can be computed using the following formula [25]:

$$\mathrm{MFCC}(l) = \frac{1}{M}\sum_{i=0}^{M-1} \log(E(i)).\cos\left(\frac{2\Pi}{M}\left(i+\frac{1}{2}\right).l\right)$$

where M represents the number of Mel filters, and $l = 0 \ldots M-1$.

The feature vector that represents the distinctive properties of the audio signals is designed to be up to 39 MFCC parameters in length, consisting of 12 Cepstral coefficients and an energy component, along with their dynamic and acceleration coefficients (Δ and $\Delta\Delta$). The acoustic vector files (.mfcc) obtained in this step will

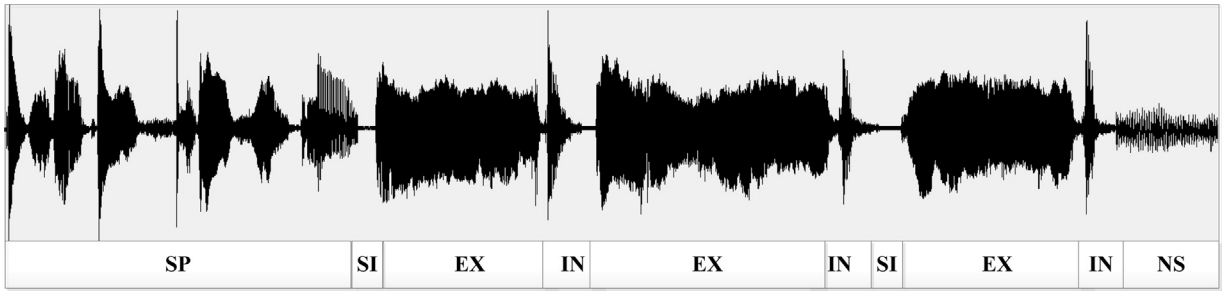


Fig. 3. Example of manual segmentation using the Wave Surfer tool.

be used in both the training and recognition phases of the system, and the extracted MFCCs will be useful for the NCDS processing phases, as no additional computation is required to extract the features. This makes our proposed NCDS a feasible and timesaving segmentation system.

5.3. Training

Our segmentation system consists of 6 classes, corresponding to the six types of sounds composing the audio recordings: Expiration (EX), Inspiration (IN), BIP sounds, Silence (SI), and Noise (NS). The main goal in this step is to establish a consistent pattern representation for each class or label, which is also called a statistical model or HMM model. An HMM model is defined at the training stage as a reference model. This is because, during testing, a direct comparison should be made between the unknown label and each HMM-trained model to determine the most probable identity of this unknown label. A training phase for modeling all the acoustic activities is therefore essential.

For comparison purposes, we used 4-, 5-, and up to 8-state HMMs, in which the first and last states are non-emitting states. Moreover, multiple Gaussian distributions, varying from 1 to 100, with diagonal matrices are used and described by mean and variance vectors. The optimal values of the HMM parameters, the transition probability, mean, and variance of each observation, are estimated iteratively at this point. This is also called re-estimation, as the procedure is repeated many times for each HMM until convergence is reached. The Baum–Welch algorithm is used to estimate and re-estimate the mean and covariance of the each model.

5.4. Recognition stage

After the training stage has been completed, a different HMM is trained for each of the six classes. At the recognition stage, the trained HMMs are used to generate a set of transcriptions for unknown observations. Therefore, given an unknown observation or unknown segment of an audio recording, the unknown observation is converted into a series of feature vectors (.mfcc), in the same way as in data training. This acoustic information, along with the reference HMM models, dictionary, and class names, are taken as input to create output in a label (.lab) file. Schematic of recognition stage is shown in Fig. 5.

Each HMM class produces a posterior probability estimate, and the HMM with the maximum probability is chosen as the most likely class. This phase is performed using the Viterbi algorithm [23].

5.5. Evaluation of system performance

The performance of the automatic segmentation system, such as that of any recognition system, is usually analyzed in terms of

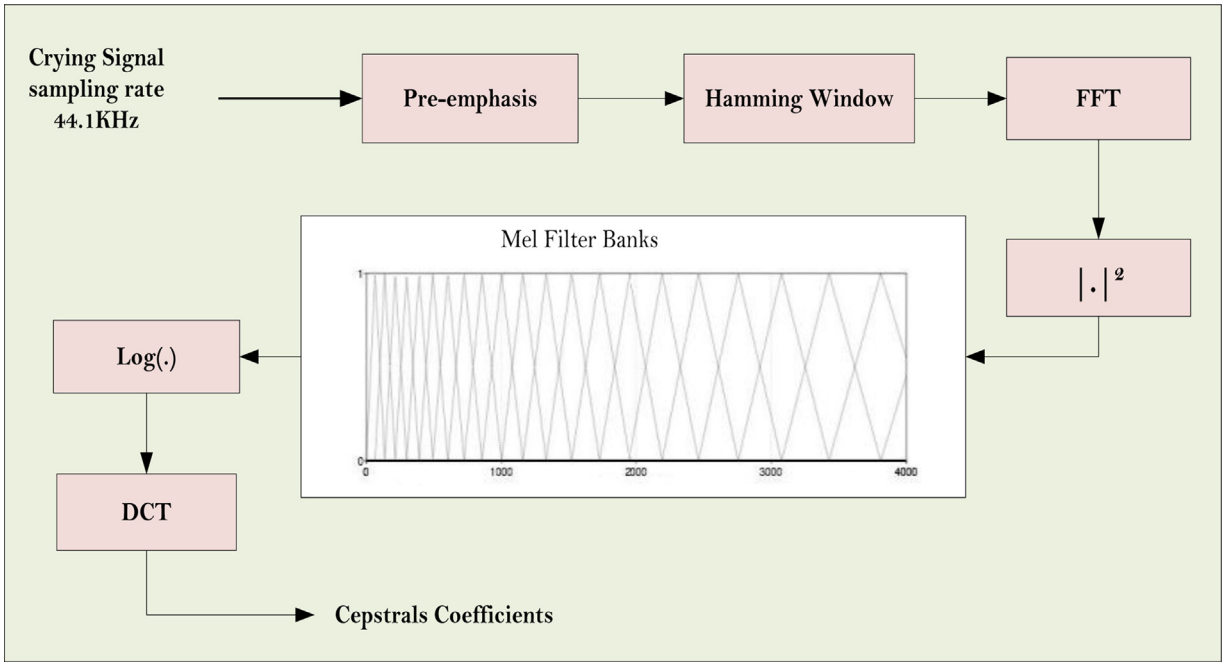


Fig. 4. Extraction Mel Frequency Cepstral Coefficients (MFCC) from the audio recording signals.

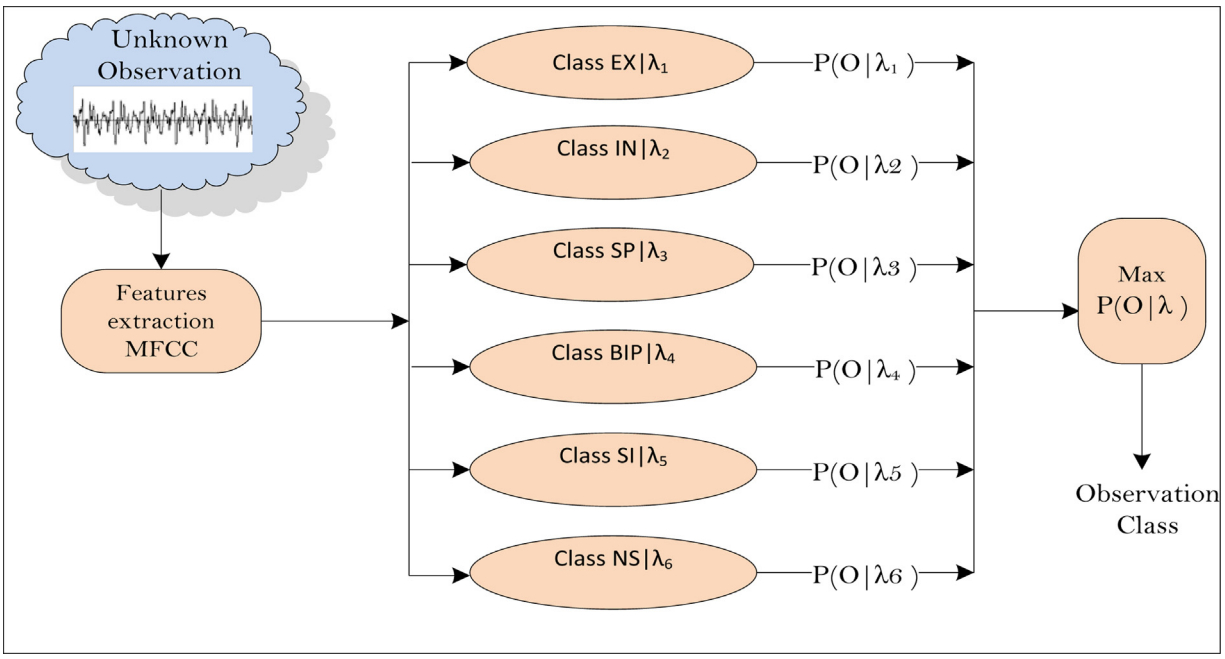


Fig. 5. Recognition stage.

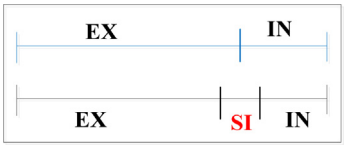


Fig. 6. Example of an insertion error – reference labels are marked in blue and automatic labeling in black – insertion error in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

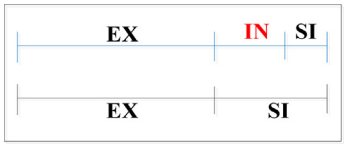


Fig. 7. Example of a deletion error – manually annotated labels in blue – automatic labeling in black and deletion error in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accuracy and speed. In this step, the output transcription file results are compared to the manually annotated files (reference labels N). Therefore, the comparison is made on a label-by-label basis by matching each of the recognized and reference label sequences.

Three possible types of errors should be calculated: number of insertions (I), deletions (D), and substitutions (S).

In Figs. 6–8, we can see the difference between the three types of errors:

Once the alignment between the automatic transcription file and the reference file is done and both of three types of errors

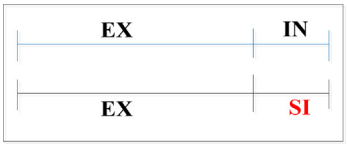


Fig. 8. Example of a substitution error – manually annotated labels in blue – automatic labeling in black and deletion error in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(*D*, *I* and *S*) mentioned above are calculated, the accuracy rate of the system could be estimated as follows: Accuracy Rate AR:

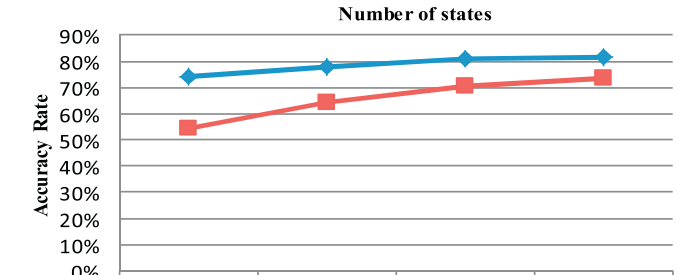
$$\mathrm{AR} = \frac{N - D - I - S}{N} \times 100\%$$

where *N* represents the total number of labels in the reference transcription files, *D* the number of deletions, *S* the number of substitutions, and *I* the number of insertions.

6. Results and discussion

The basic function of this system is to differentiate the audible expiratory and inspiratory parts of newborn cries in audio recordings. The performance of this segmentation system has been evaluated on the testing corpus using a manually segmented cry corpus. We then measured the discrepancies between manual segmentation and automatic segmentation.

Various training strategies were evaluated, to select the most suitable reference system, and some important observations were made concerning the best segmentation performance. Here, we summarize the experiments that were carried out. The first step in the manipulation of our automatic newborn cry segmentation



	4	5	6	7
EX	74.30%	78.20%	81%	81.70%
IN-EX	54.46%	64.20%	70.72%	73.87%

Fig. 9. HMM classification results obtained with the various model topologies and a fixed window size of 30 ms and MFCC_E_D_A_Z (39 parameters).

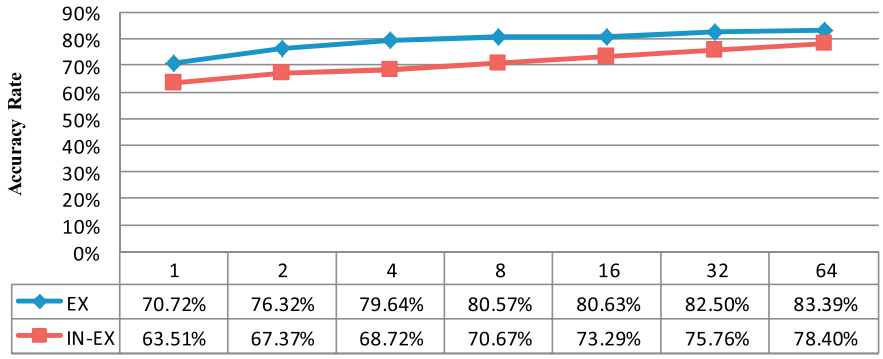


Fig. 10. Relationship between the Accuracy Error Rate and the number of mixtures for each state of a 7-state HMM with a fixed window size of 30 ms and MFCC_E_D_Z (26 parameters).

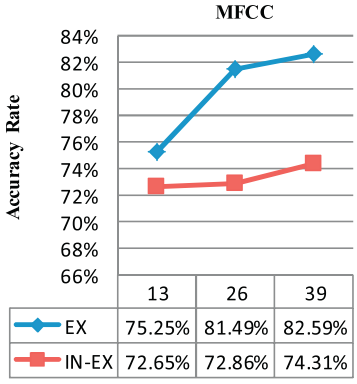


Fig. 11. Relationship between the Accuracy Error Rate and the number of MFCC parameters for a 5-state HMM and a window size of 50 ms.

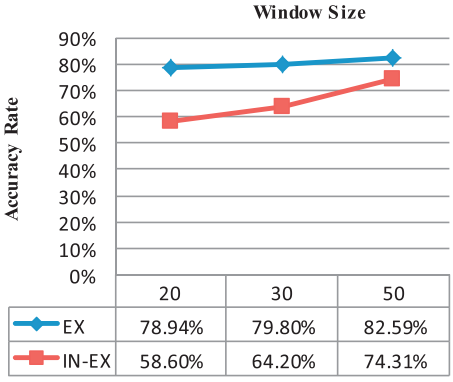


Fig. 12. Relationship between the Accuracy Error Rate and the window size using a 5-state HMM and MFCC_E_D_A_Z (39 parameters).

system is the selection of the best HMM model topology. A large number of experiments were performed to determine the optimal number of states and the number of mixtures for the training data. Figs. 9 and 10 illustrate the classification accuracy that results when these numbers are varied.

It is essential to recall here that the main goal of this research is the detection and differentiation of the expiratory and inspiratory parts of newborn cries from cry recordings. To achieve this goal, we focus now on the accuracy rate of these two classes. In Fig. 9, we show the accuracy rates obtained using HMM topologies varying from 4 to 7 states, with a fixed-length window of 30 ms with a 50% overlap and 39 dimensional feature vectors. Fig. 10 presents the accuracy results using a 7-state HMM. The observation distribution for each state is modeled by a different number of Gaussian mixtures, varying from 1 to 64. In the first two sets of experiments, we discovered that increasing the number of states and the number of mixtures has a significant impact on the system in terms of classification accuracy, returning rates of up to 83.39%. Note that the accuracy rates are higher for the expiration phase, owing to the larger number of occurrences of this type of cry in the training corpus. The expiration parts of cry sounds was the easiest to classify, which makes intuitive sense because these parts are different from the other acoustical activities recorded. The purpose of conducting the third set of experiments was to determine the effect of the number of MFCCs on the performance of the system. As Fig. 11 shows, the higher the number of MFCC parameters, the higher the classification accuracy results. These three tests were performed using the 5-state HMM, a fixed-length window of 50 ms with a 50% overlap. For comparison purposes, we also tested the impact of window size on system performance. As indicated in Fig. 12, the best rate obtained for the classification of the expiratory parts was 82.59%. We used the same 5-state HMM topology as in the previous experiment, but with a fixed number of feature vectors (39).

We looked at the way in which the performance of the proposed method varied with changes in a large number of its components. Table 3 illustrates the accuracy results obtained as a function of the HMM topology, the number of Gaussian mixtures, the number of filter banks of the 12 MFCC, and the window size. The experiments show that the best accuracy is achieved using a 7-state HMM and feature vector with 39 components and a window size of 50 ms (Best values are marked in bold in Table 3). This accuracy

Table 3
Overall system accuracy obtained for the MFCC with 39 parameters.

Number of states	MFCC_E_D_A_Z (39 parameters)					
	20 ms		30 ms		50 ms	
	INS-EXP	EXP	INS-EXP	EXP	INS-EXP	EXP
4	51.76	75.15	54.46	74.3	64.68	79.68
5	58.6	78.94	64.2	78.2	74.31	82.59
6	67.15	79.27	70.72	81	75.76	83.39
7	74.24	80.1	73.87	81.7	**77.93**	**83.79**

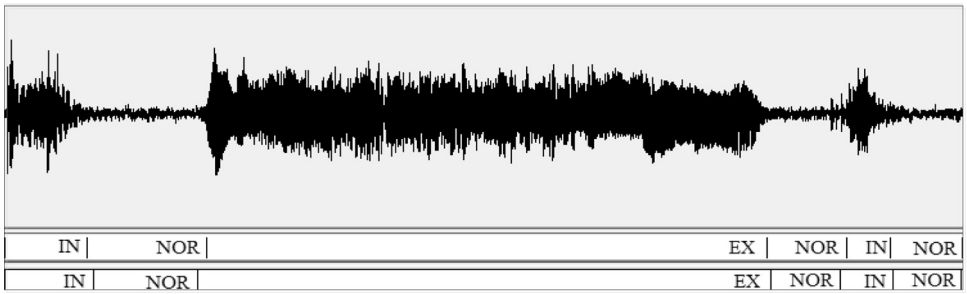


Fig. 13. Illustration of the lack of precision in boundary selection between the manual segmentation (top) and the output of the segmentation system.

is as high as 83.79% for the expiration episodes and 77.93% for the classification of both the expiration and inspiration sounds.

On the whole, the best performance is achieved with a window length of 50 ms, and using a 7-state HMM and an MFCC with 13 parameters, along with its energy component, delta, and accelerations. Typically, the approach wrongly estimates, either positively or negatively, the starting and ending points of the expiratory and inspiratory episodes. Fig. 13 illustrates a comparison between a manual segmentation (top) and the output of the proposed segmentation system. This lack of precision in boundary selection should be treated in a future work using temporal approaches.

7. Conclusions

The paper presents our preliminary results in our ongoing work to develop a tool for the automatic segmentation of newborn cry signals designed to detect the audible expiratory and inspiratory components of the newborn cry.

The authors have demonstrated the effect of some parameters on the performance of the proposed segmentation system. The difficult issues inherent in the manual labeling process are somewhat alleviated in the automated procedure. The system has two main stages, which are feature extraction and training and recognition, which both are performed using the HTK.

This research has revealed that the performance of our automatic newborn cry segmentation system is enhanced by making a number of design improvements:

- Use of an HMM topology with a larger number of states to obtain a higher accuracy rate.
- Use of multiple mixtures of Gaussian components for a more robust segmentation process.
- Use of a larger number of filter banks and MFCC parameters to achieve better trajectory modeling.
- Consideration of the impact of window size.

We applied several parameter variations to find the optimal configuration for our segmentation system. By comparing the experimental sets, we found that the best system performance is achieved using the following set of parameters: an MFCC feature vectors with 39 parameters, and an HMM of window size of 50 ms and 7 states. We conclude that the current system's performance can be considered satisfactory. The automatic segmentation system is convenient to operate and gives consistent results. In general, it could be used to create a starting point for further manual refinement if desired or a post-processing stage. The system has been used on recordings of varying types of content with an acceptable degree of success.

Our system gives an accuracy of more than 83%, however, it has some limitations owing to a lack of training data. Consequently, additional manually segmented data need to be added to the training corpus, which should improve the accuracy rate of the system. In the future, its applicability and performance are expected to continue to grow. Based on the results we have obtained, future work can be aimed at improving the system's segmentation performance, which could be achieved by better training of some unit models and adding some post-processing techniques to the system. Many more parameters should be incorporated in any future study to address the problematic issues that were noted in this study.

The tool that we have developed is also part of the effort to develop our automatic Newborn Cry-based Diagnostic System (NCDS). It will serve as the front-end processing unit for the NCDS to improve this system's recognition performance.

Acknowledgments

The authors acknowledge The Bill and Melinda Gates Foundation, which supported the Newborn Cry-based Diagnostic System project through the Grand Challenges Explorations Initiative. The authors are grateful to the staff of the Neonatology Departments at Saint-Justine Hospital in Montreal and Sahel Hospital in Lebanon for their collaboration.

References

[1] S. Orlandi, C. Manfredi, L. Bocchi, M.L. Scattoni, Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis, Conf. Proc. IEEE Eng. Med. Biol. Soc. 2012 (2012) 2953–2956.
[2] H.L. Golub, A Physioacoustic Model of the Infant Cry and Its Use for Medical Diagnosis and Prognosis, MIT Press, 1980.
[3] A. Proctor, Pathological cry, stridor and cough in infants: a clinical-acoustic study, J. Hirschberg & T. Szende, Akademiai Kiado, 1982. (156 pp., 109 Illustrations: Two 33 1/3 RPM Records, U.S. $28.00) (Distibutors: Kultura, Hungarian Foreign Trading Company, P.O.B. 149, H-1389, Budapest.), Infant Ment. Health J. 5 (1984) 245–247.
[4] G. Várallyay, Future prospects of the application of the infant cry in the medicine, Period. Polytech. 50 (April) (2005) 47–62.
[5] M.A. Rui, L.C. Altamirano, C.A. Reyes, O. Herrera, Automatic identification of qualitatives characteristics in infant cry, in: Spoken Language Technology Workshop (SLT), IEEE, 2010, pp. 442–447.
[6] H. Farsaie Alaie, C. Tadj, Cry-based classification of healthy and sick infants using adapted boosting mixture learning method for Gaussian mixture models, Model. Simul. Eng. 2012 (2012) 10.
[7] M. Hariharan, J. Saraswathy, R. Sindhu, W. Khairunizam, S. Yaacob, Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks, Expert Syst. Appl. 39 (8) (2012) 9515–9523.
[8] A. Fort, C. Manfredi, Acoustic analysis of newborn infant cry signals, Med. Eng. Phys. 20 (September) (1998) 432–442.
[9] S.M. Grau, M.P. Robb, A.T. Cacace, Acoustic correlates of inspiratory phonation during infant cry, J. Speech Hear. Res. 38 (April) (1995) 373–381.
[10] K. Michelsson, O. Michelsson, Phonation in the newborn, infant cry, Int. J. Pediatr. Otorhinolaryngol. 49 (Suppl. 1) (1999) S297–S301.
[11] K. Wermke, W. Mende, C. Manfredi, P. Bruscaglioni, Developmental aspects of infant's cry melody and formants, Med. Eng. Phys. 24 (September–October) (2002) 501–514.
[12] K. Kuo, Feature extraction and recognition of infant cries, in: 2010 IEEE International Conference on Electro/Information Technology (EIT), 2010, pp. 1–5.
[13] A. Zabidi, W. Mansor, K. Lee Yoot, R. Sahak, F.Y.A. Rahman, Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism, in: 5th International Colloquium on Signal Processing & Its Applications, 2009, CSPA, 2009, pp. 204–208.

[14] A.I.G. Várallyay Jr., Z. Benyó, The automatic segmentation of the infant cry, in: Előadás kivonatok. Méréstechnikai, Automatizálási és Informatikai Tudományos Egyesület, 2008.
[15] A.I.G. Várallyay Jr., Z. Benyó, Automatic infant cry detection, in: 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Firenze, 2009.
[16] R. Cohen, Y. Lavner, Infant cry analysis and detection, in: 2012 IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI), 2012, pp. 1–5.
[17] Y.N. Jean-Julien Aucouturier, K. Katahira, K. Okanoya, Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models, Acoust. Soc. Am. 130 (2011) 2969–2977.
[18] R.F. Orlikoff, R.J. Baken, D.H. Kraus, Acoustic and physiologic characteristics of inspiratory phonation, J. Acoust. Soc. Am. 102 (September) (1997) 1838–1845.
[19] A. Verduzco-Mendoza, E. Arch-Tirado, C.A. Reyes-Garcia, J. Leybon-Ibarra, J. Licona-Bonilla, Spectrographic cry analysis in newborns with profound hearing loss and perinatal high-risk newborns, Cir. Cir. 80 (January–February) (2012) 3–10.
[20] L.L. LaGasse, A.R. Neal, B.M. Lester, Assessment of infant cry: acoustic cry analysis and parental perception, Ment. Retard. Dev. Disabil. Res. Rev. 11 (2005) 83–93.
[21] S. Young, G. Evermann, The HTK Book, 1996.
[22] M. Gales, S. Young, The application of hidden Markov models in speech recognition, Found. Trends Signal Process. 1 (2008) 195–304.
[23] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[24] K. Sjölander, J. Beskow, Wavesurfer – an open source speech tool, in: INTERSPEECH, 2000, pp. 464–467.
[25] L.R. Rabiner, B.H. Juang, Fundamentals of Speech Recognition. Prentice-Hall signal Processing Series, 1993.