TAMPERE UNIVERSITY OF TECHNOLOGY

Gaurav Naithani

Acoustic Analysis of Infant Cry Signals

Master of Science Thesis

Examiners:
Associate Prof. Tuomas Virtanen
MSc. Katariina Mahkonen

Examiners and thesis topic approved by The Faculty of Computing and Electrical Engineering on 14th January 2015

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Signal Processing
NAITHANI, GAURAV: Acoustic Analysis of Infant Cry Signals.
Master of Science Thesis, 68 pages
May 2015
Major: Signal Processing
Examiners: Associate Prof. Tuomas Virtanen, MSc. Katariina Mahkonen
Keywords: Infant cry analysis, Audio segmentation, Fundamental frequency estimation

Crying is the first means of communication for an infant through which it expresses its physiological and psychological needs. Infant cry analysis is the investigation of infant cry vocalizations in order to extract social and communicative information about infant behavior, and diagnostic information about infant health. This thesis is part of a larger study whose objective is to analyze the acoustic properties of infant cry signals and use it for early assessment of neurological developmental issues in infants.

This thesis deals with two research problems in the context of infant cry signals: audio segmentation of cry recordings in order to extract relevant acoustic parts, and fundamental frequency ($F0$) estimation of the extracted acoustic regions. The extracted acoustic regions are relevant for extracting parameters useful for drawing correlation with developmental outcomes of the infants. Fundamental frequency ($F0$) , is one such potentially useful parameter whose variation has been found to correlate with cases of neurological insults in infants. The cry recordings are captured in realistic hospital environments under varied contexts like infant crying out of hunger, pain etc. A hidden Markov model (HMM) based audio segmentation system is proposed. The performance of the system is evaluated for different configurations of HMM states, number of component Gaussians, and using different combinations of audio features. Frame based accuracy of 88.5 % is achieved. YIN algorithm, a popular $F0$ estimation algorithm, is utilized to deal with the fundamental frequency estimation problem, and a method to discard unreliable $F0$ estimates is suggested. The statistics associated with distribution of $F0$ estimates corresponding to different components of cry signals are reported.

This work would be followed up to find meaningful correlations between extracted $F0$ estimates and developmental outcomes of the infants. Moreover, other acoustic parameters would also be investigated for the same purpose.

PREFACE

This work was has been conducted at Department of Signal Processing, Tampere University of Technology, Finland.

First of all, I would like to express my utmost gratitude to my supervisor, Dr. Tuomas Virtanen, for his constant support, invaluable guidance and immense patience throughout this work. I would also like to thank Katariina Mahkonen and Toni Heittola for providing me valuable suggestions during the initial stages of this work. I have immensely benefited from our discussions and I am thankful for your guidance in these initial steps of my scientific career.

In addition, I am grateful to Dr. Jukka Leppänen and Jaana Kivinummi, both from Infant Cognition Laboratory, University of Tampere, for giving me the opportunity to work in this project and providing the necessary funding to make this work possible.

Moreover, I wish to express my appreciation to the entire team of Audio Research Group for providing me the supportive and stimulating environment for carrying out this work. I would also like to thank my friends in Finland for making my stay away from home a cherishable one and my friends in India for their moral support.

Finally, I owe my biggest thank to my parents and my sister for supporting me through each endeavour of mine. Without their sheer support, endless love and sacrifices, I would not be where I am today.

Gaurav Naithani
Tampere, 19th May 2015

CONTENTS

TERMS AND DEFINITIONS

MFCC	Mel-frequency Cepstral Coefficient
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
HMM	Hidden Markov Model
VQ	Vector Quantization
GMM	Gaussian Mixture Model
pdf	Probability Distribution Function
EM	Expectation Maximization
MLE	Maximum Likelihood Estimate
F0	Fundamental frequency
ZCR	Zero Crossing Rate
ACF	Autocorrelation Function
SIFT	Simple Inverse Filter Tracking
ASR	Automatic Speech Recognition

List of Figures

List of Tables

1. INTRODUCTION

Infant cry analysis is a multidisciplinary field of research and researchers from various fields, e.g., pediatrics, developmental psychology, communication sciences, and signal processing, have contributed to it. Crying can perhaps be regarded as the first means of communication for an infant with its environment. It conveys to the caregiver any physiological or psychological requirement the infant may have and is therefore an important indicator of the biological and psychological status of the infant. There are two kinds of information that we can derive from cry sounds of infants: health related information, and social or psychological information. In order to extract these, acoustic analysis of cry signals and perception experiments [1] have been performed. In this thesis, our interest is in diagnostic value of infant cry.

Acoustic analysis of infant cry has existed as an active field of research for a long time. It generally involves analysis of acoustic characteristics of cry signals, e.g., fundamental frequency, temporal variation of fundamental frequency, amplitude, formants, etc. [1–9]. The primary motive that fueled research in this field was the possibility of finding correlations between extracted cry characteristics and medical condition of ailing infants. These correlations, once found, could then possibly be used for the development of a diagnostic tool. Any such diagnostic tool should have some or all of the following characteristics:

1. It would be non invasive in nature. It would help in diagnosis of conditions which can only be detected by invasive procedures.

2. It would be able to detect conditions which warrant immediate diagnosis. Sudden infant death syndrome (SIDS) is one such condition which involves sudden, unexpected death of an infant, usually during sleep. Researchers have pointed to a relation between cry characteristics and SIDS [6, 10].

3. It would be useful for prognosis of long term neurological development of the infant.

1.1 Objective of the Thesis

This thesis is a part of a larger study at University of Tampere on analyzing infant cry recordings for finding potentials markers of child development and health. The study aims to develop a method making it possible to detect health and developmental issues in infants at very early age. Infant cognition Laboratory, University of Tampere School of Medicine; and Audio Research Group, Tampere University of Technology are contributing to this study.

The scope of this thesis is divided into two tasks. The first task is to devise methods to extract relevant parts from the infant cry recordings. These recordings are captured in real hospital environment and thus contain background noises as well as portions which are not useful for further analysis. The composition of an infant cry elucidating its useful and irrelevant parts will be discussed in more detail in Section 1.2. The second task is to develop analytical tools to analyze the extracted relevant parts. Fundamental frequency ($F0$) estimation is one such analytical tool which has been widely used in infant cry research. The same is investigated in this thesis.

This study will involve further development of analytical tools and investigation of other acoustic characteristics of infant cry signals apart from $F0$. The aim would be to use them for the purpose of deriving meaningful correlations between the cry characteristics and cognitive developmental outcomes of the infants. The scope of this thesis is however limited as this analysis will not be a part of it.

1.2 Infant Cry Signal

An infant cry signal consists of a series of expirations and inspirations produced by an infant. The expirations and inspirations are separated by bouts of silence. The signal may also contain non cry vocals produced by the infant in between the series of expirations and inspirations.A cry signal captured in a realistic environment like pediatric ward of a hospital usually also contains background noise which may be contributed by the environment in which the recording is done, or by the recording equipment itself. Hence the signal can be thought of as a combination of expirations, inspirations, non cry vocals, and background noise.

The terminology used in infant cry literature to describe components of a cry signal can sometimes be quite confusing. In order to avoid this, we will now define some terms,

- An *expiratory phase* is a bout of expiration in the cry recording separated from the other bouts of expiration/ inspiration by a period of silence.
- An *inspiratory phase* is a bout of inspiration in the cry recording separated

from other bouts of expiration/ inspiration by a period of silence.

It should be noted that the above mentioned period of silence between an expiratory phase and the inspiratory phase following it can be very short in some instances. Figure 1.1 is an example of portion of a cry recording captured in hospital environment. It depicts the different components of cry signal discussed here.

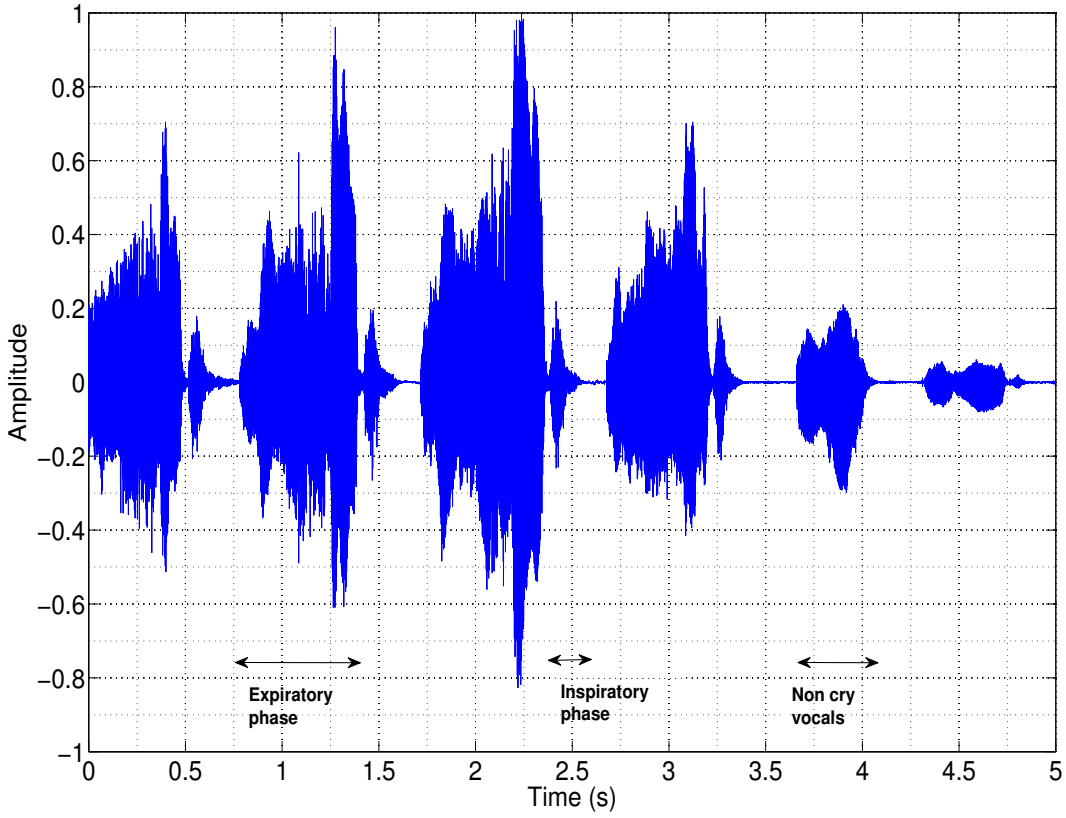

Figure 1.1: An example of infant cry signal exhibiting expiratory and inspiratory phases.

In this thesis, we aim to extract the relevant parts from cry recordings captured under realistic hospital environment in the presence of non cry vocals and background noise. Most of the previous infant cry research has been primarily focused on analysis of expiratory phases, and analysis of inspiratory phases has been given comparatively less attention. The relevance of anatomical and physiological bases of inspiratory phonation has been pointed out by Grau et al. [11]. In this thesis, we aim to develop a method to segregate expiratory phases as well as inspiratory phases from the cry recordings. The regions of interest, namely, expiratory and inspiratory phases, have to be accurately identified in the cry recordings in presence of background noise and non cry vocals. One solution to this problem is manually

annotating the recordings using some sound editing software, e.g., *Audacity*. This method is subjective, time consuming and prone to errors. Moreover, it becomes impractical if the number of audio files to be annotated is large. In such a case, there is need for an automated method to be developed. We have proposed a hidden Markov model (HMM) based audio segmentation system to achieve this. Using it, the cry signal under inspection is segregated into what we call the regions of interest,namely expiratory and inspiratory phases, and regions not significant for our purpose, which in this thesis we term as residual. Residual is basically a garbage class consisting of acoustic regions except the above mentioned regions of interest.

The regions of interest would then be used to extract parameters which would then be used for finding meaningful correlations with the developmental outcomes. Fundamental frequency $F0$ is one such crucial parameter. In this thesis, we have investigated fundamental frequency estimation of these regions of interests using YIN algorithm. YIN algorithm [12] is a popular pitch estimation algorithm used for speech and music processing.

1.3 Organization of the Thesis

The thesis is organized as follows:

Chapter 2 presents a brief review of previous work done in the field of infant cry analysis in the contexts of audio segmentation and fundamental frequency ($F0$) estimation. It is followed by a description of theoretical concepts of audio segmentation and fundamental frequency estimation which are used in this thesis.

Chapter 3 presents a description of the implemented systems proposed to solve the problems of audio segmentation and fundamental frequency ($F0$) estimation for infant cry signals. An HMM based audio segmentation system to extract expiratory and inspiratory phases from cry recordings has been proposed. Subsequently, application of YIN algorithm to infant cry signals and a method to refine the obtained $F0$ estimates is described.

Chapter 4 is devoted to evaluation of the implemented systems described in Chapter 3. The performance metric used for evaluation and the data set on which experiments were conducted is described as well.

Chapter 5 summarizes the entire work done in this thesis and suggests directions for future work.

2. THEORETICAL BACKGROUND AND LITERATURE REVIEW

The chapter serves as the theoretical background to this thesis. Two problems have been investigated in this thesis: audio segmentation of the infant cry recordings to extract the acoustic regions of interest and fundamental frequency estimation of these regions. The chapter starts with a review of the previous work done in the field of infant cry research in the context of above two research problems. It is followed by a discussion on the fundamentals of audio segmentation and fundamental frequency estimation on which the proposed solutions described in this thesis are based upon. This chapter lays the foundation for clear understanding of the implemented systems which will be described in next chapter.

2.1 Literature Review

Infant cry research, in its initial days, was based on auditory identification of various cry types [13]. In the decades of 1960 and 1970, advancement in the sound recording technology like sound spectrographs led to progress in this field. Sound spectrograms of healthy as well as sick infants were analyzed to obtain acoustic characteristics from which a number of descriptive characteristics could be derived [13–15]. This method was heavily dependent on subjective visual examination rather than quantitative objective methods and allowed for derivation of only a limited number of acoustic characteristics. The other issues plaguing it were poor dynamic range and poor frequency resolution of the sound spectrograms [16]. Moreover, this method was unsuitable for analysis in cases where a large number of audio files needed to be examined in a short period of time. Advancement in the computing technologies and signal processing methods allowed for the use of computer based methods. Using these methods, a number of useful acoustic parameters could now be derived directly instead of relying upon visual examination alone.

It has been postulated that emission of cry sounds by the infant is not mere an acoustic-linguistic event. Researchers have long been trying to extract diagnostic, communicative and predictive information contained in it. Infants suffering from specific medical conditions are known to produce cry sounds different from healthy

infants. It has also been argued that neurological status of an infant is interlinked with the cry signal it produces [1,2]. A Cry signal is produced by a complex biological phenomenon which is a combination of neural and physiological mechanisms [9]. Its correlation with medical conditions like encephalitis [17], Down's syndrome [18, 19], Cri-du-chat syndrome [20], cleft palate [8], brain damage [21, 22], etc., have been widely studied. In these studies, acoustic characteristics were mainly extracted from cry signal spectrograms and correlations were drawn with the associated medical conditions of the infants.

During the days of spectrographic analysis, the audio segmentation problem was solved through visual inspection of sound spectrograms. Voiced crying sounds were manually selected from the spectrograms of the cry recordings [2]. With the advent of computer assisted methods for processing audio signals, it became possible to extract specific regions from the cry recordings, which would then be utilized for extraction of useful acoustic parameters. In many research efforts, the problem of audio segmentation has been addressed as problem of *voicing determination* and the problem of $F0$ estimation has been framed as being the preceding or subsequent stage to it. *Voicing determination* is the problem of labeling each audio region under consideration as either voiced or unvoiced. It is the voiced audio regions that contribute to $F0$ estimation [23, 24]. In such cases, the audio signal is either pre-processed to determine regions of interest (voiced audio regions) beforehand or post-processed to extract regions having meaningful F0 (voiced audio regions). Various audio segmentation approaches have been reported apart from the traditional approach of manual segmentation [25]. Use of commercial or freely available software [5, 26] has been quite popular as well. Várallyay et al. [3] have used modified harmonic product spectrum (HPS) based methods to extract expiratory phases from the recordings while treating inspiratory phases as noise. Aucouturier et al. [27] have previously used HMMs for segmenting cry recordings in a way similar to what we have attempted in this thesis. We have investigated different configuration of HMM states and experimented with some additional acoustic features in addition to conventional mel-frequency cepstral coefficients (MFCCs) used in [27]. In most of these studies inspiratory phases have not been treated as a separate class and the main emphasis has been on extraction of expiratory phases.

Fundamental frequency of an infant cry signal corresponds to the rate of glottal opening and closing in the vocal tract. The larynx, also known as voice box, houses the vocal chords of an infant and is responsible for generating fundamental frequency of a cry signal. It is postulated that larynx of an infant is controlled by the cranial nerves of the nervous system [28, 29]. Moreover, fundamental frequency is found to be exhibiting higher levels and more variability in infants suffering from neurological insults [1].

Fundamental frequency estimation is a complex task and the research in this field has been largely context dependent. Hence, an $F0$ estimation algorithm has to be chosen depending upon the context in which it is expected to perform. There have not been many attempts of developing $F0$ estimation algorithms specifically for cry signals. Use of commercial or freely available softwares has been quite popular for $F0$ estimation and *voicing determination* in infant cry research. Two of the most widely used systems are Praat [30] and Computerized Speech Laboratory (CSL) [31]. Praat has been used by Baeck et al. [32], Esposito et al. [4], Lin et al. [33] and Irwin [34]. Similarly, CSL speech lab [31] has been used by Wermke et al. [7], Rautava et al. [35] and Mampe et al. [36]. Others have utilized $F0$ estimation algorithms devised for speech and music signal processing. Simplified inverse filter tracking (SIFT) algorithm [37] and its modifications have been used by Kheddache et al. [26], Lederman [38] and Manfredi et al. [24]. Similarly, Várallyay et al. [3] employed smooth spectrum method (SSM) for $F0$ estimation, and cepstrum analysis has been utilized by Reggiannini et al. [23].

2.2 Audio Segmentation: Feature Extraction

Audio segmentation is an important preprocessing method in audio signal processing. The objective of audio segmentation is to divide an input audio signal into acoustically homogeneous regions/classes. The output is a labeled audio signal on which further analysis can be selectively performed on the region/class of choice depending upon the application. There are two ways of segmenting an audio signal into the regions of interest. It can either be done via unsupervised classification or via supervised classification. In this thesis we have employed the latter. It consists of two steps:

1. Feature extraction
2. Pattern recognition

Feature extraction involves converting a raw cry signal into a sequence of acoustic feature vectors carrying characteristic information about the signal. The most popular choice for features in the field of audio and speech processing is mel frequency cepstral coefficients (MFCCs) [39]. The same has been used as the principal feature vector for cry signals in this thesis. In addition to MFCCs some other features have been also been experimented with. They will be described in Section 2.2.3.

2.2.1 MFCC

Mel frequency cepstral coefficients (MFCCs) are inspired by psychoacoustic model of human auditory perception. The human ear does not interpret frequency in a linear manner. The information carried by low-frequency components is more important than carried by high frequency components [40]. Moreover, it can not resolve between frequencies lying within the same critical band [41] and this effect becomes more pronounced at higher frequencies. Mel scale, which is a perceptually motivated scale of frequencies, exploits this property of human auditory system. It arranges the frequencies in such a way so that the frequencies perceived by the listener are equal in distance from each other. Mel scale is approximately linear below 1 kHz and logarithmic above it. It can be derived from the linear frequency scale using the mathematical expression

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700}), \tag{2.1}$$

where f is frequency on the linear frequency scale in Hz. In order to extract MFCCs, an audio signal is first broken down into short time frames, e.g., 25 ms with 50 percent overlap and then multiplied with a window function. This is followed by computation of the fast Fourier transform (FFT) for each frame. The phase information is subsequently discarded and magnitudes are squared to obtain the power spectrum. This power spectrum is subjected to frequency warping from the linear frequency scale to mel-frequency scale using the mel-scale triangular filterbank. As the frequencies get higher, the width of the filters also increases. This reflects the fact that ability of the human ear to resolve closely spaced frequencies decreases as the frequency increases. The spectral energies within a band are then summed up to obtain filterbank energies which are then subjected to logarithm operation. Finally, the discrete cosine transform (DCT) is computed of the log filterbank energies according to the equation

$$c(i) = \sqrt{\frac{2}{N_f}} \sum_{j=1}^{N_f} e_j \cos(\frac{\pi i}{N_f}(j - 0.5)), \tag{2.2}$$

where $c(i)$ is the i^{th} MFCC coefficient, e_j is logarithm of the energy of the j^{th} filter in the filter bank for $j = 1, 2, ...N_f$, and N_f is the number of mel filters in the filterbank. Again, both the steps, summing the energies within a band and the taking logarithm of filterbank energies, are inspired by the model of human auditory perception. Generally 13 MFCC coefficients are extracted for each frame. Figure 2.1 illustrates the whole process of MFCC extraction.

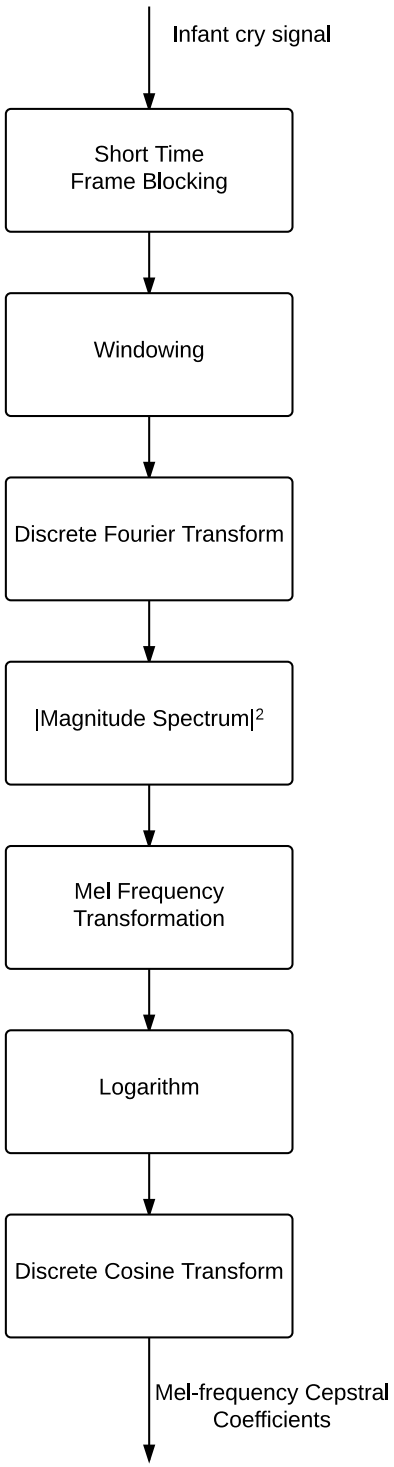

Figure 2.1: Extraction of Mel frequency cepstral coefficients.

2.2.2 Delta and Delta-Delta Features

The MFCC feature vector computed as described in Section 2.2.1 contains the information of only the power spectral envelope of a signal frame, but it fails to capture the temporal dynamics of the audio signal. Delta features are used to capture this

dynamics. They are basically time derivative of the MFCC features. Delta-delta features are in turn time derivatives of the delta features and similarly capture the temporal dynamics of . Delta and delta-delta features are also referred as differential and acceleration coefficients, respectively. They have been widely used in the field of speech recognition, where generally they are used in conjunction with MFCC feature vectors. Delta coefficients are calculated from MFFCs as

$$\mathbf{del}_n = \frac{\sum_{l=1}^{L} l(\mathbf{c}_{n+l} - \mathbf{c}_{n-l})}{2\sum_{l=1}^{L} l^2}, \tag{2.3}$$

where $\mathbf{c}_n$ is the MFCC vector corresponding to n^{th} signal frame. MFCC vectors for frames ranging from $n - L$ to $n + L$ are utilized to compute delta coefficient vector $\mathbf{del}_n$ for n^{th} frame, L being the window size. Delta-delta coefficients are similarly calculated using delta coefficients in the place of MFCCs. These features are concatenated with the static MFCC features to give a combined feature matrix. The delta and delta-delta features have rarely been used in the field of infant cry analysis.

2.2.3 Other Features

The aim of audio segmentation in this thesis is to successfully discriminate between expiratory and inspiratory phases. The acoustic characteristics that help in this objective may prove to be useful features. In addition to the standard features used in speech and audio signal processing, i.e., MFCCs, deltas and delta-deltas, we have experimented with several other features, namely

1. *Fundamental frequency*: Fundamental frequency of a quasi-periodic signal, e.g., infant cry, has been defined in Section 2.4.1. Expiratory and inspiratory phases are known to have different distributions of fundamental frequencies [11] with inspiratory phases exhibiting higher means and standard deviations. Hence, this property can be utilized for achieving our audio segmentation objectives.

2. *Aperiodicity*: Aperiodicity is the measure of harmonicity of the signal frame. Expiratory phases are generally more harmonic than inspiratory phases. This difference in harmonicity can be exploited via the aperiodicity feature. It has been defined in more detail in Section 2.4.4.

3. *Running averages and running variances*: A moving average filter can be employed on the MFCCs to calculate running average vector $\bar{\mathbf{u}}$. Similarly, a

moving average filter can be employed on the square of MFCCs to get vector $\bar{\mathbf{u}}_2$. Running variances can then be calculated using $\bar{\mathbf{u}}$ and $\bar{\mathbf{u}}_2$ using

$$\mathbf{u_{var}} = \bar{\mathbf{u}}_2 - (\bar{\mathbf{u}})^2, \tag{2.4}$$

where $\mathbf{u_{var}}$ is the running variance vector, and the second term of the above equation is simply the square of each component of running average vector $\bar{\mathbf{u}}$, each component here representing the running average of corresponding MFFC term. Note that Equation (2.4) is employed upon a window of length W_l signal frames. In order to compute $\mathbf{u_{var}}$ corresponding to a particular time frame, both $\bar{\mathbf{u}}$ and $\bar{\mathbf{u}}_2$ are computed within the window of size W_l centered at that frame. For the subsequent time frames, the window is shifted accordingly and features are computed. Different window sizes, W_l can be employed to calculate these features.

2.3 Audio Segmentation: Pattern Recognition

The output of feature extraction stage is a sequence of feature vectors denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n, ..., \mathbf{x}_Z\}$, where subscript n denotes the frame index and Z denotes the total number of frames in the signal. In the second step, this sequence of extracted features vectors is fed to a pattern classifier to get an output class label for each of the Z frames. Segment boundaries for different class segments in the signal can be deducted from these output labels. Statistical models like hidden Markov models (HMMs) have been quite popular in conventional speech processing for pattern classification. HMMs have been successfully applied to the problem of speech recognition [42] to model variability in speech caused by different speakers, speaking styles, vocabularies, and environments. In this thesis, HMMs have been been used in the context of infant cry signals to model the variation of expiratory and inspiratory phases in the cry signals from different infant subjects and recorded under varying environmental conditions. In this section, hidden Markov models will be formally defined, and the associated terms will be explained.

2.3.1 Distrete Time Markov Chains

Discrete time Markov chain is a stochastic process which takes on a finite number of states from a set of N possible states, $S = \{s_1, s_2, ..., s_N\}$. At each time instant[1] $t = 1, 2, ...$, the system undergoes a transition from state s_i to state s_j. This state transition is governed by a probability a_{ij}, known as the state transition probability.

[1]In this section, Markov chains are described in general and hence a notation of system transitions with respect to time index t is adopted. In the context of this thesis, we have an audio system where system transitions would occur each frame index n. The two notations are thus equivalent.

The system may make a transition from a particular state into a different or into the same state. Let the states of the system at time instants t and $t-1$ be q_t and q_{t-1}, respectively. In general, the probability that the system is in state q_t, is a function of complete history of the system which makes the analysis quite complicated. Here "the Markov property" can be utilized to simplify the analysis. The Markov property asserts the principle, "Given the present, the future is independent of the past," which essentially means that the system is memoryless. Hence the probability of the system to be in state q_t depends only upon the preceding state q_{t-1} and not on the entire past history of the states taken by the system. Mathematically, this can be expressed as

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_{t-2}, ..., q_1 = s_1) = P(q_t = s_j | q_{t-1} = s_i) \quad . \tag{2.5}$$

The corresponding joint probability for a sequence of Z states, $(q_1, q_2, q_3, ..., q_Z)$ is given by

$$P(q_1, q_2, ..., q_Z) = \prod_{z=1}^{Z} P(q_z | q_1, ..., q_{z-1}) = P(q_1) \prod_{z=2}^{Z} P(q_z | q_{z-1}) \quad . \tag{2.6}$$

This is known as the first order Markov assumption, which implies that the memory of the system is restricted to one preceding state. Similarly, $\mathfrak{n}^{th}$ order Markov chain can be constructed which is able to memorise $\mathfrak{n}$ preceding states instead of just one. The state transition probabilities, a_{ij}, are assumed to be stationary, i.e., they are independent of time. Mathematically, this be expressed as

$$a_{ij} = P(q_t = s_j | q_{t-1} = s_i), \qquad 1 \leq i, j \leq N \quad . \tag{2.7}$$

Moreover, in accordance with the laws of probability, we have

$$a_{ij} \geq 0 \ \text{ and } \ \sum_{j=1}^{N} a_{ij} = 1 \quad . \tag{2.8}$$

The state transition probabilities can be represented in the form of a state transition probability matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix} \tag{2.9}$$

where the i^{th} row represents the probability of transitioning from i^{th} state to all other possible states. In order to completely characterize a Markov model, probability of a state of being the initial state of the system has to be defined. This is called the initial state distribution. Let π_i be the probability that i is the initial state of the system then

$$\mathbf{\Pi} = \{\pi_{s_1}, \pi_{s_2}, ..., \pi_{s_N}\} = \{P(q_1 = s_1), P(q_1 = s_2), ..., P(q_1 = s_N)\} \tag{2.10}$$

constitutes the initial state distribution of the system, where

$$\sum_{i=1}^{N} \pi_i = 1 \quad . \tag{2.11}$$

A Markov model is said to be ergodic if it is possible to reach one state from all the other states in a finite number of steps. Figure 2.2 illustrates an example of such a model.

2.3.2 Hidden Markov Models

A discrete time Markov model assumes that each state of the system can be uniquely associated with an observable event. This assumption is too restrictive and it holds true for only simple modeling tasks. Discrete time hidden Markov models are an extension of discrete time Markov models where a state is no longer associated with a single output observation, but is capable of generating a number of outputs according to a probability distribution. The name, "hidden Markov model", implies that the sequence of states taken by the system is not directly observable and can not be deduced with absolute certainty by observing the outputs. The actual underlying states of the system are therefore "hidden".

Rabiner and Juang [43] define an HMM as, "A doubly stochastic process with an underlying stochastic process that is not observable, but can only be observed through another set of stochastic processes that produce the sequence of observed symbols". Here the first stochastic layer is a first-order Markov process consisting

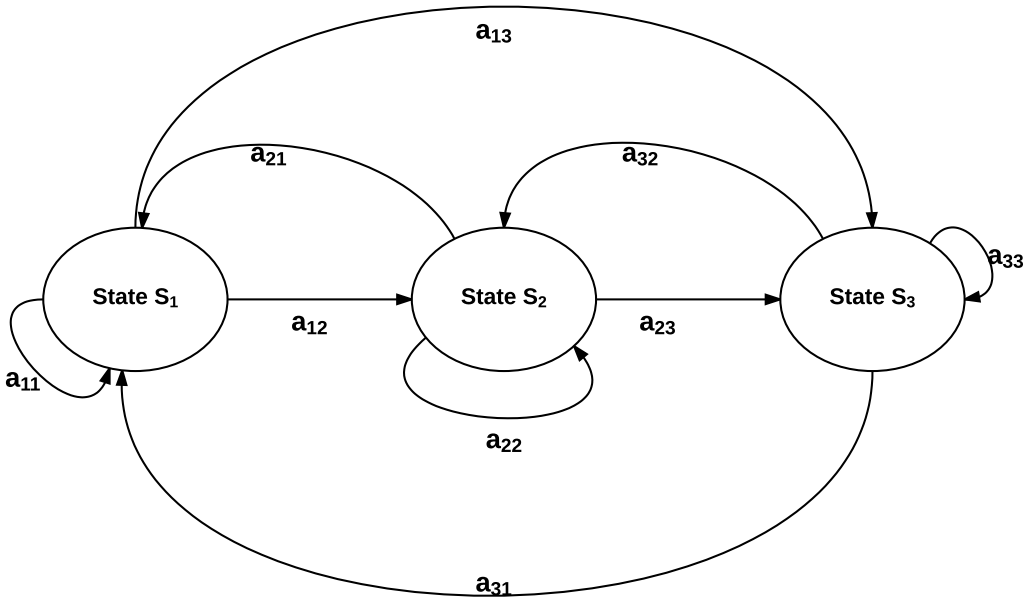

Figure 2.2: An example of an ergodic Markov model.

of hidden states characterized by a set of initial state probabilities Π, and state transition probabilities A. The second stochastic layer produces observable outputs, $V = \{v_1, v_2....., v_K\}$, for the hidden states, $S = \{s_1, s_2, ..., s_N\}$. If we represent the probability of observing output v_k at time instant t when the underlying state of the system is s_j as $b_j(k)$, then the set of all such probabilities for K observable outputs and N hidden states can be denoted as

$$\begin{aligned} B_s &= \{b_j(k)\} \\ \text{where} \quad b_j(k) &= P(o_t = v_k | q_t = s_j) \ . \end{aligned} \tag{2.12}$$

o_t is the observation emitted at time instant t. This distribution is known as the state emission distribution. In accordance with the laws of probability, we have

$$b_i(k) \geq 0 \ \text{ and } \ \sum_{k=1}^{K} b_i(k) = 1 \quad . \tag{2.13}$$

Each underlying state thus has a probability distribution over the set of possible output observations. For N underlying states and K possible output observations, 2.12 can be expressed as a $N \times K$ matrix

$$\mathbf{B} = \begin{pmatrix} b_1(1) & b_1(2) & \cdots & b_1(K) \\ b_2(1) & b_2(2) & \cdots & b_2(K) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(1) & b_N(2) & \cdots & b_N(K) \end{pmatrix} \tag{2.14}$$

which is known as the emission probability matrix of the system. Figure 2.3 illustrates an example of a hidden Markov model producing a sequence of observations and having three underlying states. Note that in such a model, state transitions are possible either only in the forward direction or to the same state from which it originated. Such a model is known as left-to-right HMM.

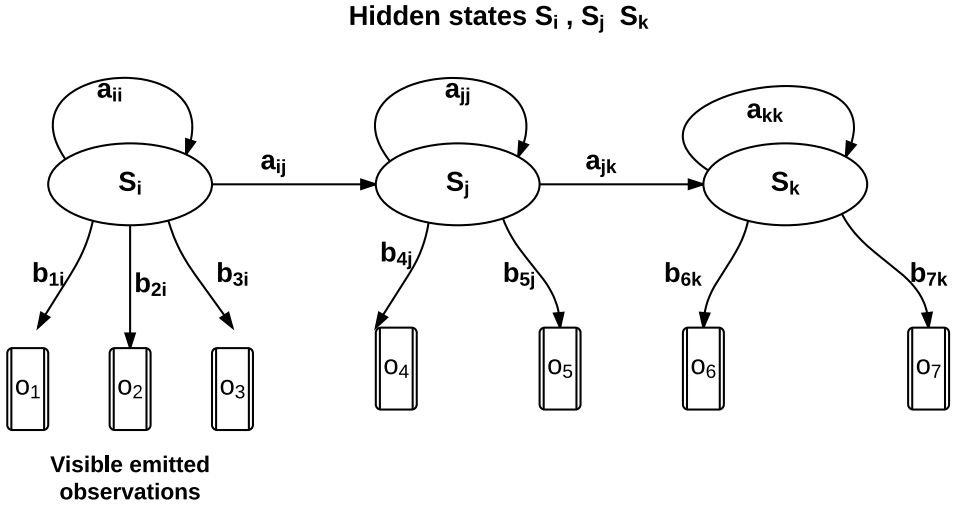

Figure 2.3: Left to right HMM topology.

In order to completely characterize a hidden Markov model, we need the following components [42]:

1. N, the number of underlying hidden states $S = \{s_1, s_2, ..., s_N\}$ taken up by the system.

2. K, the number of discrete output states $V = \{v_1, v_2, ...v_K\}$ generated by the sequence of hidden states S, which are observable.

3. A set of initial state probabilities $\mathbf{\Pi} = \{\pi_i\}_{i=1}^{N}$ given by Equation (2.10).

4. A set of state transition probabilities $\mathbf{A} = \{a_{ij}\}$ given by Equation (2.9).

5. A set of state emission probabilities $\mathbf{B} = \{b_j(k)\}$ given by Equation (2.14).

The notation $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$ is often used in literature as a compact way to represent HMMs. In addition to the Markov assumption, the following properties are assumed in order to make the HMMs mathematically and computationally tractable,

1. *Stationarity assumption*: The state transition probabilities are assumed to be independent of the time t at which the actual state transition takes place, i.e., Equation (2.7) holds for all values of t.

2. *Output independence assumption*: The emitted output observations $V = \{v_1, v_2, ..., v_K\}$ are conditionally independent of each other, i.e., the probability of observing $o_t = v_k$ at time t is independent of previous observations $o_{t-1}, o_{t-2}, ..., o_1$, and the underlying states $q_{t-1}, q_{t-2}, ..., q_1$, given the current state q_t.

On the basis of the method of modeling state emission probabilities, HMMs can be divided into three different categories, namely

1. *Discrete HMMs*: The output observation sequence V consists of discrete outputs. Each underlying state has a probability mass function which for all states can be represented in the form of Equation (2.14). In the context of speech recognition, the output observations correspond to quantization levels of a vector quantization (VQ) codebook [44]. Discrete HMMs offer the advantage of reduced computation, although systems based on them are less flexible and suffer from inaccuracies due to quantization errors [45].

2. *Continuous HMMs*: The output observation is a continuous variable instead of a discrete variable. The outputs are generally modeled by a mixture of probability density functions. In order to ensure that the model parameters can be re-estimated in a consistent manner, some restrictions are applied to the form of the observation probability density function (pdf) [42]. A mixture of Gaussian probability density functions, i.e., Gaussian mixture model (GMM), is chosen as the most common form of representation for the output observation. For continuous HMM case, Equation (2.12) can be expressed as

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm}\, \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \mathbf{\Sigma}_{jm}) \tag{2.15}$$

where $\mathbf{o}_t$ is the vector being modeled for the system at time t and j^{th} state s_j, c_{jm} is the mixture coefficient for the m^{th} mixture component, M is the number of Gaussian pdfs in the mixture and $\mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$ is the m^{th} Gaussian distribution with mean vector $\boldsymbol{\mu}_{jm}$ and covariance matrix $\boldsymbol{\Sigma}_{jm}$. The m^{th} component for multivariate Gaussian distribution is thus given by the expression,

$$\mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) = \frac{1}{(2\pi)^{\frac{D}{2}} \ |\boldsymbol{\Sigma}_{jm}|^{\frac{1}{2}}} \ \exp(-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jm})^{\mathcal{T}} \ \boldsymbol{\Sigma}_{jm}^{-1} \ (\mathbf{o}_t - \boldsymbol{\mu}_{jm})) \quad . \tag{2.16}$$

where exp denotes the exponential function, $\boldsymbol{\Sigma}_{jm}^{-1}$ denotes the inverse of covariance matrix $\boldsymbol{\Sigma}_{jm}$ and $\mathcal{T}$ denotes the matrix transpose operation. Moreover, the mixture weights c_{jm} follow the stochastic constraints

$$\begin{gathered} \sum_{m=1}^{M} c_{jm} = 1 \quad \forall j \in \{1, 2, ..., N\} \\ \text{and, } \ c_{jm} \geq 0 \quad \forall j \in \{1, 2, ..., N\}, \ m \in \{1, 2, ..., M\} \end{gathered} \tag{2.17}$$

Continuous HMMs, while on one hand avoid some of the shortcomings of the discrete HMMs like quantization errors, on the the other hand, require considerable large amount of training data and training times [45].

3. *Semi continuous HMMs*: It is the combination of the above two HMM types. Similar to the Discrete HMMs, it involves use of vector quantization, but each VQ codeword is regarded as a continuous pdf. It is similar to parameter tying of a continuous HMM such that the states share the same distribution, which in effect happens to be the VQ codebook [46]. They have become less popular due to improvements in the estimation techniques for more efficient models like continuous HMMs, and availability of sufficient amount of training data for training these efficient models.

In order to apply an HMM to the real world problems, we must be able to:

1. Evaluate the probability of an observation sequence $\mathbf{O} = (o_1, o_2, ...,)$ if we know the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$. In other words, evaluate $P(\mathbf{O}|\lambda)$, the likelihood of the observation given the model. This is known as *probability evaluation* problem. It can be used to score several competing models and choose the best one out of them. *Forward algorithm* and *Backward algorithm* are used to solve this problem.

2. Finding out the sequence of underlying states $\mathbf{Q} = (q_1, q_2, ...)$ that best explains a given sequence of observations $\mathbf{O} = (o_1, o_2, ...)$ if we know the model

parameters $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$. This is known as the *decoding* problem and is solved by a sequential decoding algorithm known as the *Viterbi algorithm.*

3. Estimating the parameters of the model $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$ in order to maximize the probability of observing an observation sequence $\mathbf{O} = (o_1, o_2, ..)$, i.e., maximize $P(\mathbf{O}|\lambda)$. In other words, finding parameters of the model which best fits a given sequence of observations. This is known as the *training* problem for an HMM and is solved by *Baum-Welch algorithm.*

In this thesis, we are concerned with the latter two problems.

Baum-Welch algorithm: The first problem of our interest is estimating the parameters of an HMM given a set of observations consisting of features extracted from a cry signal. Estimating HMM parameters involves estimating transition probabilities, initial state distribution, and emission probability distribution from training data. In the context of this thesis, it would involve the estimation of state transition probabilities and parameters associated with Gaussian mixture models (GMM) which constitute the emission probability distribution in the continuous HMM case. The GMM parameters to be estimated are mixture weights, means, and variances of the component Gaussian distributions. The initial state distributions are computed by assuming each one of the states to be the initial state with equal probability.

Solving the training problem amounts to choosing an HMM from a set of possible models which best explains the given observations. Probabilistically, it can be framed as the problem of maximizing probability of an HMM model given the observations, i.e., $P(\lambda|\mathbf{O})$, which can be framed as *maximum likelihood estimation* problem

$$\lambda_{opt} = \arg\max_{\lambda} P(\mathbf{O}|\lambda) \tag{2.18}$$

where λ_{opt} is the optimal model that best fits the observations $\mathbf{O}$. Equation 2.18 is very difficult to solve analytically [43]; hence, an iterative approach has to be adopted. Baum-Welch algorithm [47] is a form of *Expectation Maximization* (EM) algorithm [48] which iteratively refines the model parameters until Equation 2.18 is maximized. An EM algorithm iteratively alternates between two stages : expectation (E)-step and maximization (M)-step. We start with a random estimate of HMM parameters λ, which can be computed from prior information if available. The E step estimates likelihood of observations under current parameters, and M step then uses the computed likelihoods to re-estimate the model parameters. This allows the estimate of model parameters to be refined in each step of the iterative procedure until no further improvement in the likelihood function is achieved. It also implies

that the final likelihood obtained is only a local maximum and can not be guaranteed to be a global maximum. A detailed mathematical treatment of the algorithm can be found from [42]. In this thesis, AHTO toolbox from Audio Research Group, Tampere University of Technology, has been used to train the HMMs.

Viterbi algorithm: The second problem of interest in this thesis is finding an optimal sequence of states given a trained HMM model and observation sequence. The observation sequence is the set of acoustic features vectors derived from a cry signal. This can be solved by Viterbi algorithm [49] which was invented by Andrew Viterbi as a solution to the problem of decoding convolutional codes. It is basically a dynamic programming algorithm which . The search space mentioned here is the set of all possible combinations of hidden states. The algorithm maximizes the probability of occurrence of state sequence $\mathbf{Q} = (q_1, q_2, ...)$ while observing the observation sequence $\mathbf{O} = (o_1, o_2, ...)$ when we already know the model parameters λ. Mathematically, it can be put as

$$S_{opt} = \arg\max_{\mathbf{Q}} P(\mathbf{Q}|\mathbf{O}, \lambda) \tag{2.19}$$

where S_{opt} is the optimal sequence of states. In practice, log probabilities are maximized instead. The search space can be formulated as a trellis graph structure. At each time step t, all paths leading to a particular state from all possible old states at $t-1$ time step are explored, and only the one with maximum log probability is retained. This is done for all possible states at t, and corresponding log probabilities and states are saved. This procedure is followed for each time step and the path with maximum log probability is selected. The best path is then given by the one giving maximum cumulative log probability, and it is discovered by tracing back through the trellis from the state at final time step to the state at initial time step using backpointers. The backpointers mentioned here are the states of maximum log probabilities saved earlier for each time step. The whole process is analogous to breadth first search through the trellis structure with the aim of maximizing the cost, and cumulative log probabilities serving as the cost function. A detailed mathematical treatment of the algorithm can be found from [42]. Figure 2.4 shows a trellis structure depicting four possible states and four time instants for illustration. Note that at each time step, one possible paths is retained for each state yielding the maximum cumulative log probability. Finally, the path yielding overall maximum cumulative log probability is chosen.

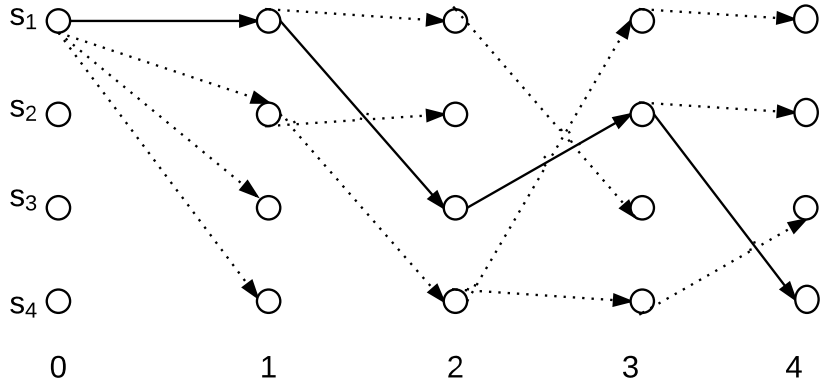

Figure 2.4: Illustration of Viterbi decoding through a search space of four states. The final optimum path is shown by dark arrows giving the output sequence, $(s_1, s_1, s_3, s_2, s_4)$. The dotted arrows show the most optimal path chosen for each state at every time step.

2.4 Fundamental Frequency Estimation

Fundamental frequency is an acoustic characteristic associated with harmonic signals. In this section, we will describe the concept of periodicity for a signal which forms the basis of the definition of fundamental frequency. It will be followed by an overview of popular fundamental frequency estimation methods. Finally, description of a popular time domain fundamental frequency algorithm called YIN algorithm will be given. YIN algorithm has been used in this thesis to solve the problem of fundamental frequency estimation in the context of cry signals.

2.4.1 Periodicity of a Signal

A signal is said to be periodic if it repeats itself at specific intervals of time. This specific interval of time is said to be the period of the signal. Mathematically, this property of a signal $y(t)$ can be described as

$$y(t) = y(t + T_0) \quad \forall t \tag{2.20}$$

where the signal $y(t)$ is a function of time t and T_0 is period of the signal. It is also evident that the above equation will hold for all integer multiples of T_0. Cheveigné and Kawahara define fundamental period as the smallest positive member of an infinite set of time shifts that leave the signal invariant [12].

Fundamental frequency, F0 is thus defined as the inverse of this fundamental period T_0,

$$F0 = \frac{1}{T_0} \, . \tag{2.21}$$

Integer multiples of fundamental frequency are referred to as harmonic frequencies. The Equation (2.20) only holds for signals which are perfectly periodic. The real world signals like speech and cry signals are of finite duration and are not perfectly periodic. Moreover, these signals often exhibit variations in periodicity with time and hence the notion of a fixed $F0$ for them is irrelevant. Such signals may however be assumed to be periodic in very small time frame in which the signal is assumed to be stationary. This is known as the assumption of quasi-periodicity. Fundamental frequency estimation is the problem of assigning a fundamental frequency for each such frame. For biological audio signals like speech and cry, fundamental frequency ($F0$) is the rate of vibration of the vocal folds of the speaker [50, 51]. $F0$ exhibits a temporal variation which is dependent upon the size and tension in the vocal folds [50].

Depending upon the domain of operation, there exists two approaches to solve the problem of fundamental frequency estimation, namely

1. *Time domain approach*
2. *Frequency domain approach*

We will now give an brief overview of different techniques which utilize these two approaches.

2.4.2 Time Domain Approach

1. *Time event rate detection*: The time event rate detection methods of $F0$ estimation rely on the principle that for a periodic signal there must be time repeating events in the signal which can be counted [52]. This information can be used to detect $F0$. *Zero crossing rate* (ZCR) [53], which involves counting how many times a signal crosses zero per unit time; *Peak rate*, which involves counting positive peaks in a signal per unit time; *Slope event rate*, which involves counting the number of zeros or peaks of the slope of a signal per unit time, are some of the time event detection methods. These methods, although fairly simple to implement, suffer from a major drawback that harmonically complex signals may have more than one events per cycle [52].

2. *Autocorrelation*: Correlation of two signals is the measure of similarity between them. It is expressed as a function of time lag τ, where τ is the lag introduced in one of the signals while keeping the other unaffected. Autocorrelation is defined as the correlation of a signal with itself. It can be expressed using the equation

$$r_t(\tau) = \sum_{j=t+1}^{t+W} y(j)\, y(j+\tau), \tag{2.22}$$

where $r_t(\tau)$ is the autocorrelation function (ACF) of signal $y(t)$, calculated at time index t with integration window size W. Note that the above expression for autocorrelation is short-time autocorrelation function calculated over frame of size W. For periodic signals, ACF exhibits peaks at zero lag as well as at lags corresponding to multiple of the fundamental period [52]. The first peak after zero thus gives the fundamental period of the signal.

The autocorrelation method gives good results for perfectly periodic signals. However, for a quasi-periodic signal consisting of multiple harmonic components, ACF peaks may correspond to the period of the constituent harmonics. Hence, a distinction has to be made between these erroneous peaks and the actual peaks corresponding to period of the overall signal. Also ACF method has tendency to pick up the formant frequency instead of the fundamental frequency [54]. Various improvements have been suggested in the ACF method to avoid these shortcomings. YIN algorithm is one such time domain method which improves upon it. It will be separately discussed in detail in Section 2.4.4.

2.4.3 Frequency Domain Approach

The frequency domain approach involves analysis of short term Fourier transform of a signal. It relies upon the principle that a periodic signal exhibits peaks in its frequency spectrum at frequencies which are multiples of fundamental frequency.

1. *Harmonic Product Spectrum*: The harmonic product spectrum method [55,56] is based on the principle that downsampling a signal by a factor of two makes the peak at second harmonic frequency in the frequency spectrum to appear at the fundamental frequency. Similarly, a downsampling by a factor of three would make the third harmonic frequency to appear at the fundamental frequency. Multiplying a signal with its various downsampled versions would make the peak at fundamental frequency to be emphasized and hence easy to extract. This methods fails in the case where harmonic component being

multiplied with the peak at fundamental frequency has very low energy. The product would be almost zero in this scenario. Moreover, the frequency resolution is dependent upon the length of FFT used which if increased would decrease the temporal resolution.

2. *Cepstral analysis*: Cepstral analysis has been used in speech signal processing to deconvolve the source excitation of speech signal and the transfer function of vocal tract of the speaker [57]. The cepstrum is formally defined as the inverse discrete Fourier transform of logarithm of the discrete Fourier transform of the signal. Mathematically, it can be expressed as

$$C(q) = \mathcal{F}^{-1}\{\log|\mathcal{F}(y(t))|\}, \tag{2.23}$$

where $C(q)$ is the cepstrum of signal $y(t)$, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote discrete Fourier transform and inverse discrete Fourier transform of the signal, respectively. The variable q here has dimensions of time and is referred to as *quefrency*. Note that the Equation (2.23) gives an expression for the real cepstrum of signal $y(t)$ as it only takes the magnitude of the Fourier transform of the signal into consideration. Taking the log magnitude of Fourier transform of the signal allows for compression of the dynamic range of equally spaced harmonic peaks in the spectrum. It essentially translates the amplitude to a usable scale. The distance between periodic harmonic peaks can be extracted in the cepstrum as a strong peak which gives the fundamental period of the signal. This methods fails for signals which do not exhibit regularly spaced harmonic partials in their frequency spectrum.

2.4.4 YIN Algorithm

YIN algorithm [12] was developed by Alain de Cheveigné and Hideki Kawahara. The algorithm derives its name from the concept of "yin" and "yang" from oriental philosophy which describes the phenomenon of contrary forces existing in a state of duality and complementing each other. In this context, these dual forces are autocorrelation and cancellation. YIN algorithm tries to overcome the shortcomings of conventional autocorrelation method, which happens to be the first step of the algorithm. The overall algorithm can be summarized in the following steps,

1. *Autocorrelation*: ACF is calculated according to Equation (2.22).

2. *Difference function*: Improvement in the ACF method is achieved via difference function. It is defined as the sum of the squares of the differences between a signal and its delayed version with time lag τ over the analysis window W. Mathematically, it is given by

$$d_t(\tau) = \sum_{j=1}^{W} (y(j) - y(j+\tau))^2, \tag{2.24}$$

where $d_t(\tau)$ denotes the difference function for the signal $y(t)$. The smallest value of time lag τ which gives a zero value for the *difference function* is the fundamental period of the signal. Here, instead of maximizing the product of the signal and its delayed version as in autocorrelation method, difference of the two is being minimized. An improvement in the error rates is achieved which can be explained by the fact that ACF is sensitive to variations in the signal. An increase in signal amplitude with time causes ACF peaks to grow with lag which in turn encourages the selection of an erroneous peak [12]. The difference function is immune to this issue as it is less sensitive to amplitude changes.

3. *Cumulative mean difference function*: The difference function calculated in step 2 is prone to picking zero lag as imperfect periodicity of the signal may force it to have non zero values at the fundamental period. One way to avoid this is to set a lower limit on the lag search range. This lower limit must also be robust against erroneous minimas of the difference function which may occur as a result of the presence of a strong first formant in the vicinity of $F0$ [12]. But the ranges of first formant and $F0$ are known to overlap [12]; hence, setting a lower limit on the search range is not a viable solution. The difference function is adapted in the form of a cumulative difference mean function in order to avoid these errors. The cumulative difference mean function $d'_t(\tau)$ is given by the expression,

$$d'_t(\tau) = \begin{cases} 1 & \text{if } \tau{=}0 \\ \dfrac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{t} d_t(j)} & \text{otherwise} \end{cases} \tag{2.25}$$

for difference function $d_t(j)$. Note that unlike $d_t(j)$ which has value 0 at zero lag, function $d'_t(\tau)$ has value 1 for zero lag. Additionally, $d'_t(\tau)$ tends to remain large at low lags, and drops below 1 only when $d_t(j)$ falls below average [12].

4. *Absolute threshold*: This step sets a threshold in order to prevent the algorithm from choosing an erroneous higher-order minima of the cumulative mean difference function given by Equation (2.25). The threshold determines a set of lags from which the smallest value of lag which gives a minima deeper than the threshold is chosen. A global minimum is chosen if none is found. The minimum can be interpreted as proportion of aperiodic power in the signal [12].

This proportion is referred to as aperiodicity in this thesis.

5. *Parabolic interpolation*: In cases where the fundamental period is not a multiple of used window length, there may be an error in the period estimate. Parabolic interpolation of the local minima of function $d'_t(\tau)$ is done in order to achieve this [12]. The interpolated minima is then used in selection of period.

6. *Best local estimate*: The last step of YIN algorithm concerns with the selection of best possible estimate in the vicinity of each analysis point. The interval $[t - 0.5T_{max}, t + 0.5T_{max}]$ is searched for minimum of $d'_\theta(T_\theta)$, where T_θ is estimate at θ and T_{max} is largest expected period [12].

Steps 5 and 6 are used to refine $F0$ estimates obtained through Step 3. Figure 2.5 depicts a chunk of a cry signal along with the computed difference function $d_t(\tau)$, and cumulative mean normalized difference function $d'_t(\tau)$. The period of the signal is determined by extracting the first minima of $d'_t(\tau)$ below the threshold which in the given example happens to be 0.3. The extracted lag value is 110 and the corresponding $F0$ value is 436.3 Hz.

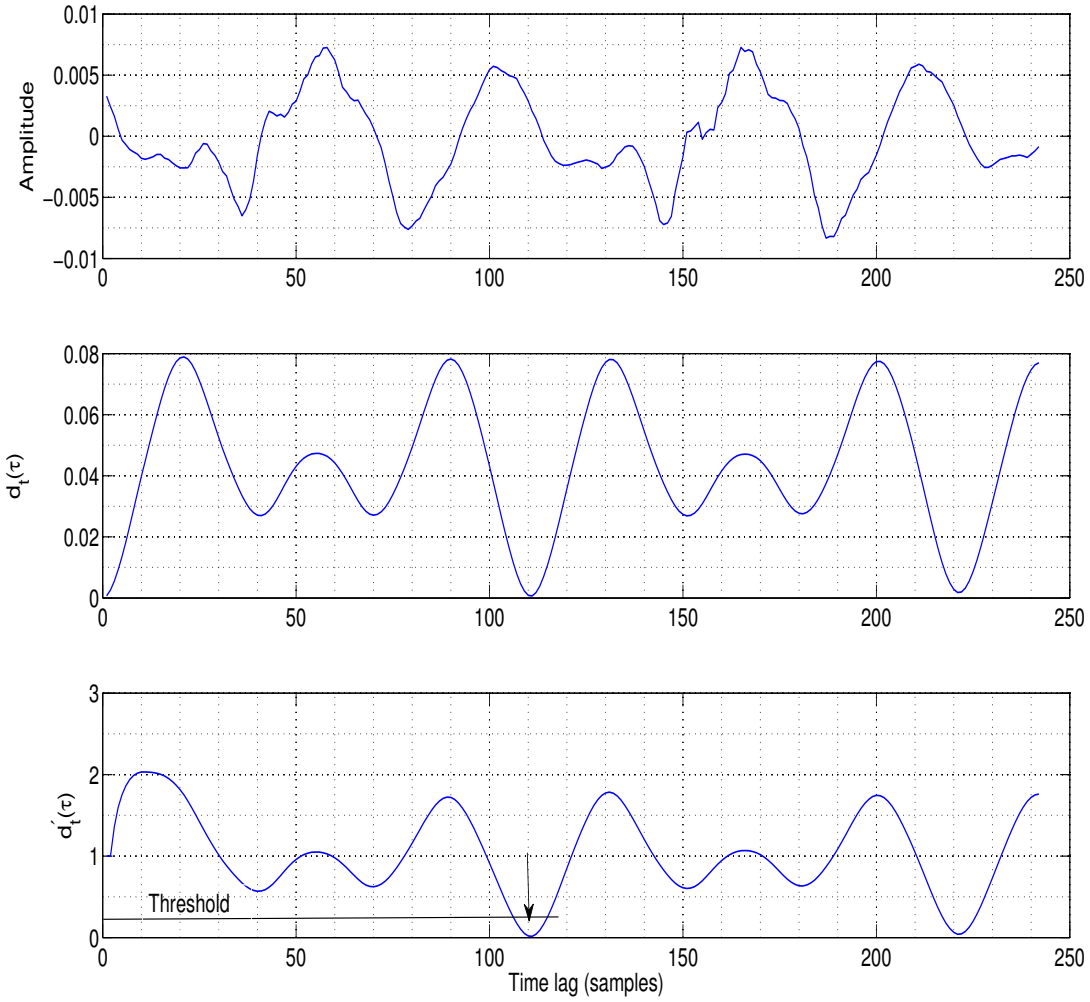

Figure 2.5: $F0$ estimation using YIN algorithm. The top panel depicts a chunk of a cry signal under investigation. The corresponding difference function $d_t(\tau)$ and cumulative mean difference function $d'_t(\tau)$ are depicted in middle and bottom panels, respectively. The value of *absolute threshold* is 0.3. Note that the first minima below threshold occurs for lag value 110 which corresponds to the fundamental period of the signal.

3. IMPLEMENTED SYSTEM

This chapter is composed of two sections. In the first section, we will describe the audio segmentation system which processes raw infant cry signals and identifies the regions of interest, i.e., expiratory and inspiratory phases. It is followed by a description of methods employed to estimate fundamental frequencies of these identified regions. The extracted fundamental frequency is one of the crucial parameters which would help in further analysis in order to achieve our ultimate goal of finding correlations between the the cry signals and developmental outcomes of the infants.

3.1 Audio Segmentation

This section describes the overall implementation of the audio segmentation system. The problem statement is segmentation of an infant cry signal into three classes: expiratory phases, inspiratory phases, and residual, by selecting appropriate features and using a pattern recognizer which in this thesis is an HMM. As explained in Section 2.2.1, the cry signal is divided into overlapping short time frames. For each short time frame, the HMM pattern recognizer gives a set of observation probabilities of the three classes being active in that frame. These probabilities are decoded using Viterbi algorithm. Figure 3.1 depicts the block diagram of the process:

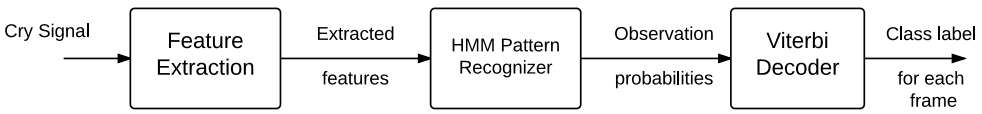

Figure 3.1: Block diagram of the audio segmentation system.

The overall implementation can be described in the following steps:

1. *Manual annotations*: In order to train our pattern recognizer and subsequently test its performance, we need labeled audio signals. Labels for the raw cry recordings can be generated by manually annotating the recordings using any audio editing software. In this thesis, we have used *Audacity* [**?**] application. Figure 3.2 is a snapshot of *Audacity* application showing a chunk of the labeled cry signal. Expiratory and inspiratory phases were annotated as depicted in Figure 3.2, and rest of the recording was considered as residual class.

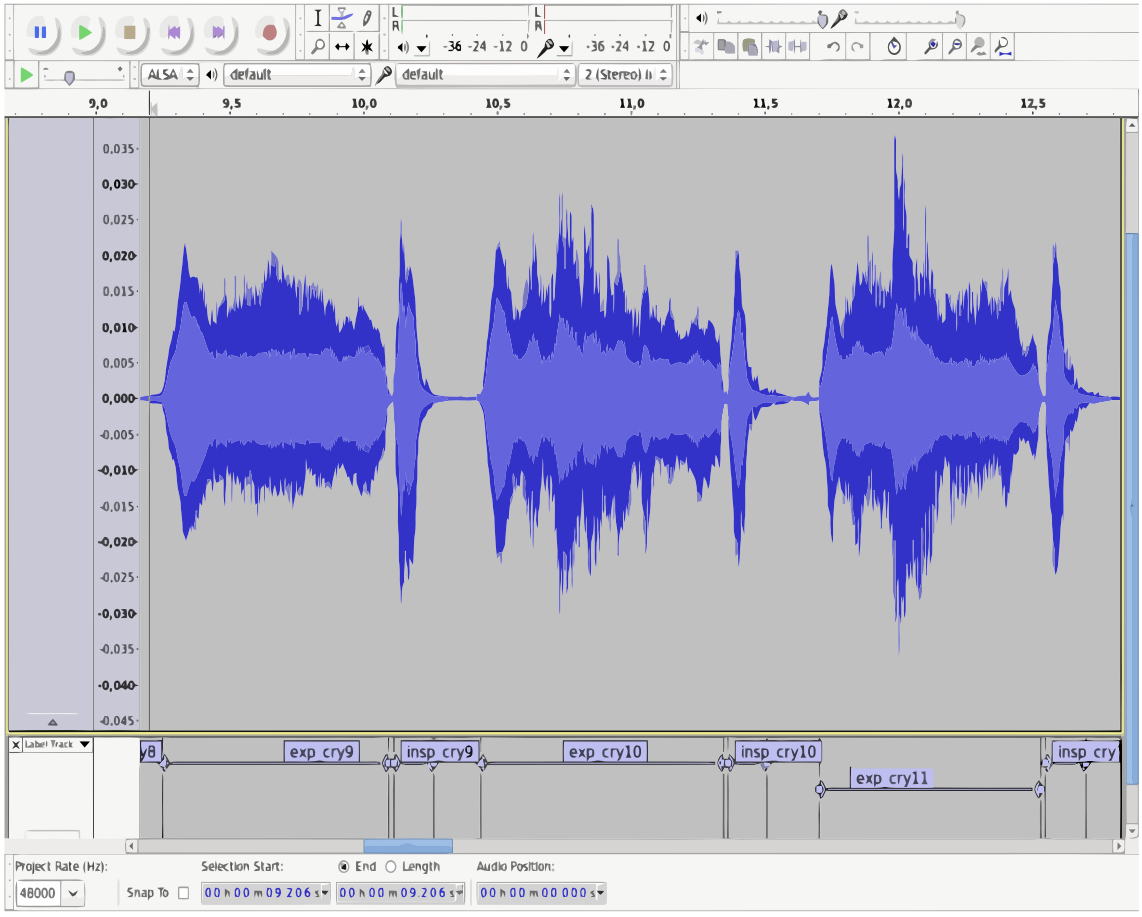

Figure 3.2: Snapshot of *Audacity* application showing a manually annotated chunk of a cry recording. Expiratory and inspiratory phases are coded by names *exp_cry* and *insp_cry*, respectively.

2. *Division of data*: A pattern recognition system uses two non overlapping data sets: a training data set which is employed for training the system, and a test data set which is used for evaluating the performance of the system. In order to generate these two data sets, we split the total available annotated data into training and test sets. The training set is 70 % of the total data set and the rest 30 % is assigned to the test set. This splitting is done on the basis of cry codes assigned to the recordings which correspond to the chronological order in which the recordings were captured. This will be explained in more detail in Section 4.2.1.

3. *Feature extraction*: Features are extracted from the manually annotated training data set. MFFCs, which are the primary audio features used in this thesis, are extracted for each 25 ms time frame having 50% overlap between the frames. For each frame, a 13 dimensional feature vector $\mathbf{x} = [x_1, x_2....., x_{13}]^T$ is extracted. This includes the zeroth order MFCC coefficient which is sum of log energies from each mel filterbank and can be thought of as a measure of energy of the frame. In conjunction with MFCCs, some additional features are also experimented with, and different combination of these features are investigated. While using these additional features, the corresponding feature vector of dimension z is appended to the MFCC feature vector. The following features are used,

   - *Deltas and delta-deltas*: There are 13 delta coefficients and 13 delta-delta coefficients for each frame, hence z here is 26.
   - *Running average and running variances of MFCCs*: There are 13 running average values and 13 running variance values for each frame, hence z here is 26.
   - *Fundamental frequency of each frame*: One fundamental frequency value is obtained for each signal frame, hence z here is 1.
   - *Aperiodicity of each frame*: One aperiodicity value is obtained for each signal frame, hence z here is 1.

   For N_i such frames for a particular target class, we have a $((13 + z) \times N_i)$ dimensional training feature matrix. Here N_i is variable with i denoting the class index. It is equal to the number of frames available in the training data set for that particular class. Features are extracted from all training files for each of the three target classes. These features extracted from different audio files are concatenated to give three training matrices, each corresponding to one of the three classes: expiratory phases, inspiratory phases, and residual class.

4. *HMM training*: Three separate HMM models are trained corresponding to the three target classes. Fully connected HMMs are used, and the standard *Baum-Welch algorithm* is used for training the models. AHTO toolbox of the Audio Research Group, Tampere University of Technology, has been used for training and testing the HMM models. HMMs being used are continuous density output HMMs for which two parameters have to be chosen: N_s, the number of states used to adequately model the class; and N_c, the number of Gaussian components in GMM used to model each state of the HMM. The effect of both these parameters on model performance has been investigated in this thesis and will be discussed in Section 4.2.3. Let us denote the number of states in the three HMMs by N_{s1}, N_{s2} and N_{s3}. Similarly, the number of component Gaussians be denoted by N_{c1}, N_{c2} and N_{c3}.

5. *Combining HMM models*: The HMMs trained for the three target classes are then combined to form a single big network model having a combined state space and transition probability matrix. The probability of transition from one model to another is determined by calculating model priors from the training data. This is done simply by counting the occurrences of that particular class from the annotated data. 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, and no term of the combined transition probability matrix is zero. There are two parameters which govern the transitions from one HMM model to another in the combined network, namely *grammar scaling factor* and *inter model transition penalty*. These parameters are widely used for tuning automatic speech recognition (ASR) systems.

    - *Grammar scaling factor* controls the weighing between acoustic and language model scores in ASR systems. In this thesis, we are not using any language model hence this factor is taken as 1 for all experiments.
    - *Inter model transition penalty* is essentially logarithmic transition penalty which controls transition from one component HMM model to another in the combined network. A more negative value leads to fewer inter model transitions and comparatively stable model. A less negative value, on the other hand, leads to a model which is frequently fluctuating among the the three target classes. We observed that a frequently fluctuating model performs better in detecting the inspiratory phases, while a more stable model is better suited for detecting expiratory phases. Thus, a trade off is required to achieve an optimal performance. Figure 3.6 depicts the predicted class labels for inter models transition penalty values -50, -20 and -1. In this thesis, we have used value of -1 which corresponds to inter

model transition probability of e^{-1}.

HMMs trained for the three target classes having N_{s1}, N_{s2}, and N_{s3} states; and N_{c1}, N_{c2}, and N_{c3} component Gaussians, respectively, can be combined to form a big HMM network which can be represented by Figure 3.3. Note that the combined model has its own transition probability matrix having dimension equal to $(N_{s1} + N_{s2} + N_{s3}) \times (N_{s1} + N_{s2} + N_{s3})$.

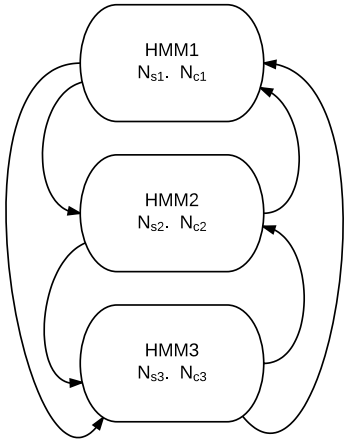

Figure 3.3: Block diagram of combined HMM model.

Following is an example of state transition matrix for the combined model with number of states in the individual HMMs, N_{s1}, N_{s2} and N_{s3} be 2, 1, and 3, respectively.

$$\mathbf{A} = \begin{pmatrix} 0.949 & 0.035 & 0.001 & 0.005 & 0.005 & 0.005 \\ 0.039 & 0.945 & 0.001 & 0.005 & 0.005 & 0.005 \\ 0.004 & 0.004 & 0.962 & 0.010 & 0.010 & 0.010 \\ 0.002 & 0.002 & 0.001 & 0.956 & 0.003 & 0.038 \\ 0.002 & 0.002 & 0.001 & 0.003 & 0.935 & 0.059 \\ 0.002 & 0.002 & 0.001 & 0.044 & 0.049 & 0.904 \end{pmatrix}$$

Where $\mathbf{A}(i, j)$ is the probability of transition from state i to state j. In the above matrix, the upper two rows with blue color corresponds to the 2 HMM states for expiratory phase class, the third row with light blue color corresponds to 1 HMM state for inspiratory phase class and the rest three rows with green

color correspond to 3 HMM states for residual class. The parameters learned for the combined HMM model are used in step 7 for decoding the observation probabilities for each test file.

6. *Generation of test observation probabilities*: Features are extracted from the test data in the same way as was done for the training data in step 3. Each test file gives a $((13 + z) \times N_i)$ feature matrix, where variable z depends on the additional feature being used, and variable N_i depends upon the number of frames in the test data file. This test feature matrix is fed to the three HMM models trained in step 4, and observation probabilities are calculated for each class. The output is a $(N_{sk} \times N_i)$ dimensional observation probability matrix consisting of probabilities for the N_i frames corresponding to each of the N_{sk} HMM states, where $k = 1, 2, 3$. Three such observation probability matrices having dimensions $(N_{s1} \times N_i)$, $(N_{s2} \times N_i)$ and $(N_{s3} \times N_i)$ are generated corresponding to three output classes.

7. *Viterbi decoding*: The observation probability matrices generated in step 6 are then vertically concatenated to give a combined observation probability matrix of dimension $((N_{s1} + N_{s2} + N_{s3}) \times N_i)$. Viterbi decoding is employed upon this combined observation probability matrix using the parameters of the combined HMM model of step 5 to give a sequence of output labels. The output is a class assignment for each frame of the test data.

Steps 4- 7 can be represented by a block diagram given by Figure 3.4.

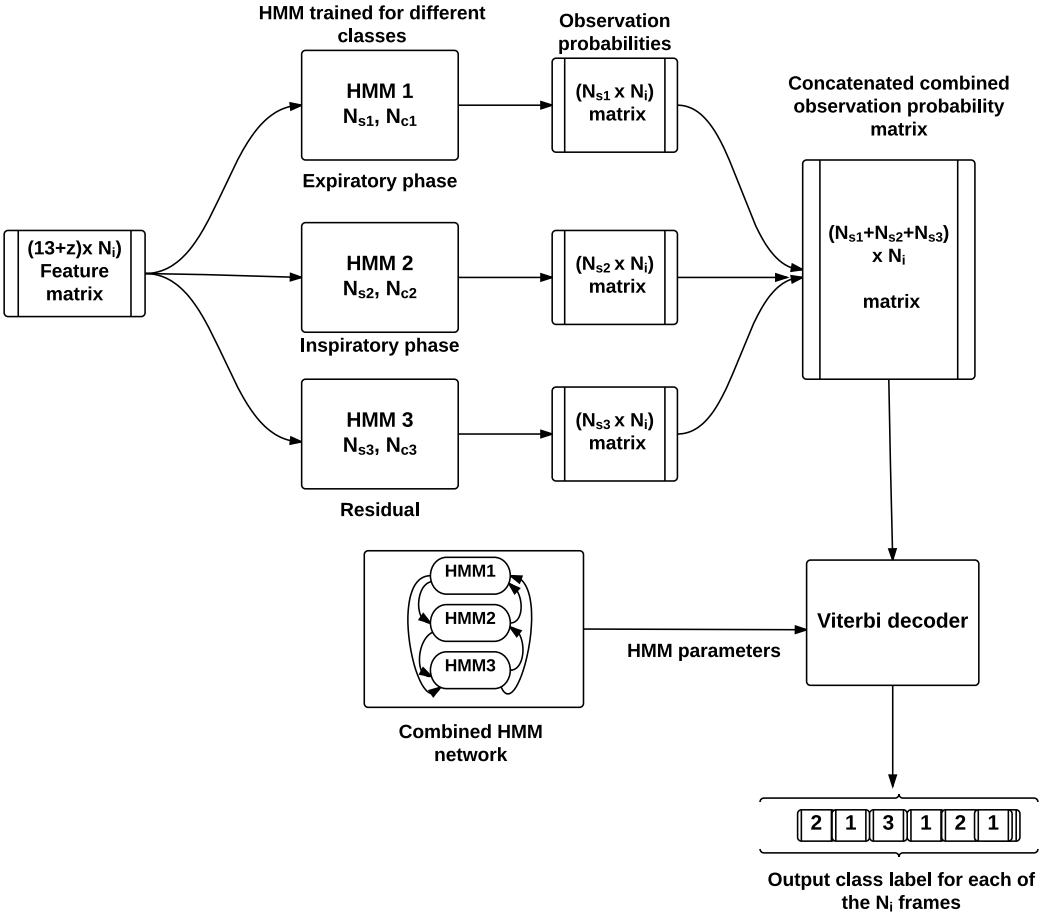

Figure 3.4: Block diagram of the audio segmentation system depicting the implementation Steps 4- 7. The input to the system is $(13 + z) \times N_i$ dimensional feature matrix derived from a test data file, where variables z and N_i depend upon dimensions of the additional features being used, and number of frames in the test data file, respectively. The output is class label assigned for each of the N_i frames.

Let us assign class labels 1, 2, and 3 to expiratory phase, inspiratory phase and residual classes, respectively. Figure 3.5 exhibits the results of segmentation for a 5 second portion of a cry signal.

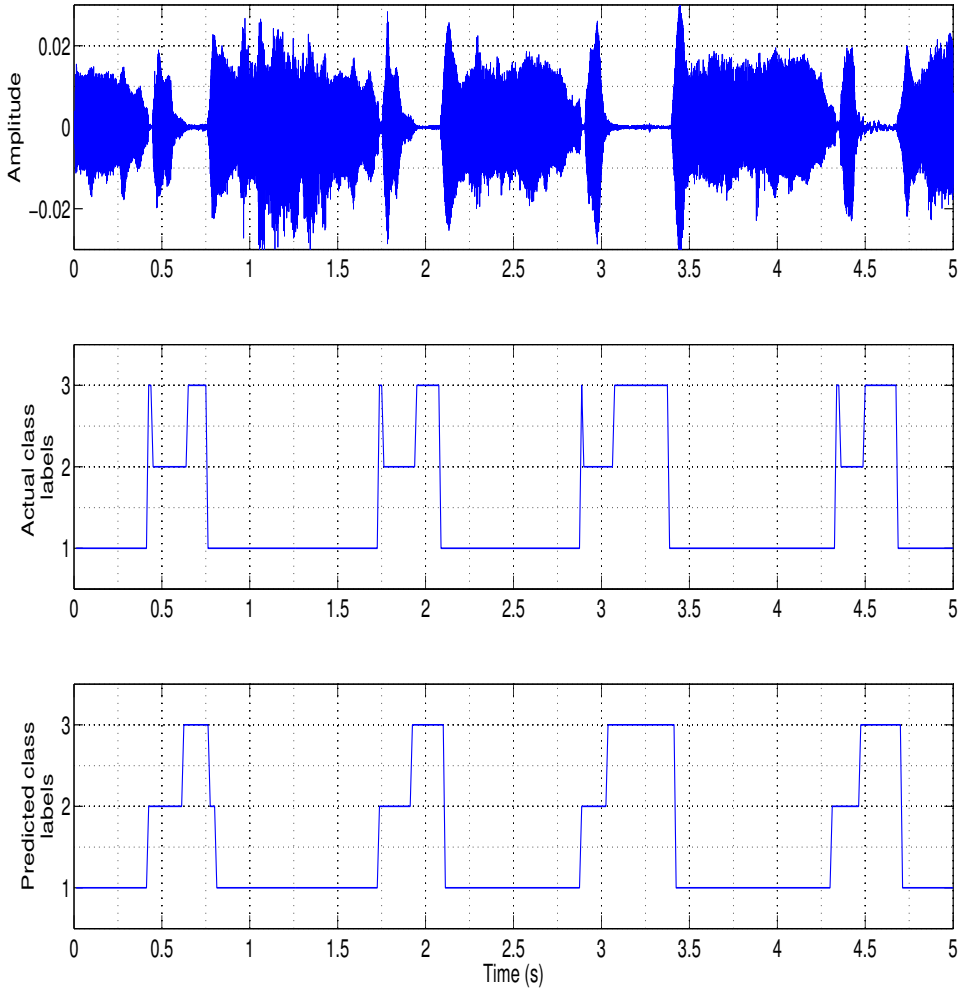

Figure 3.5: Segmentation results for a chunk of a cry signal. The class labels for expiratory phase, inspiratory phase and residual are 1, 2 and 3, respectively.

Labels generated by the HMM model are tested against the available ground truth to calculate the performance metrics of the model for the test file under inspection. Steps 6 and 7 are repeated for all the test data files and performance metrics are calculated for them. The evaluation of the model will be discussed in detail in Chapter 4 along with the used performance metrics. Figure 3.7 depicts the overall implementation.

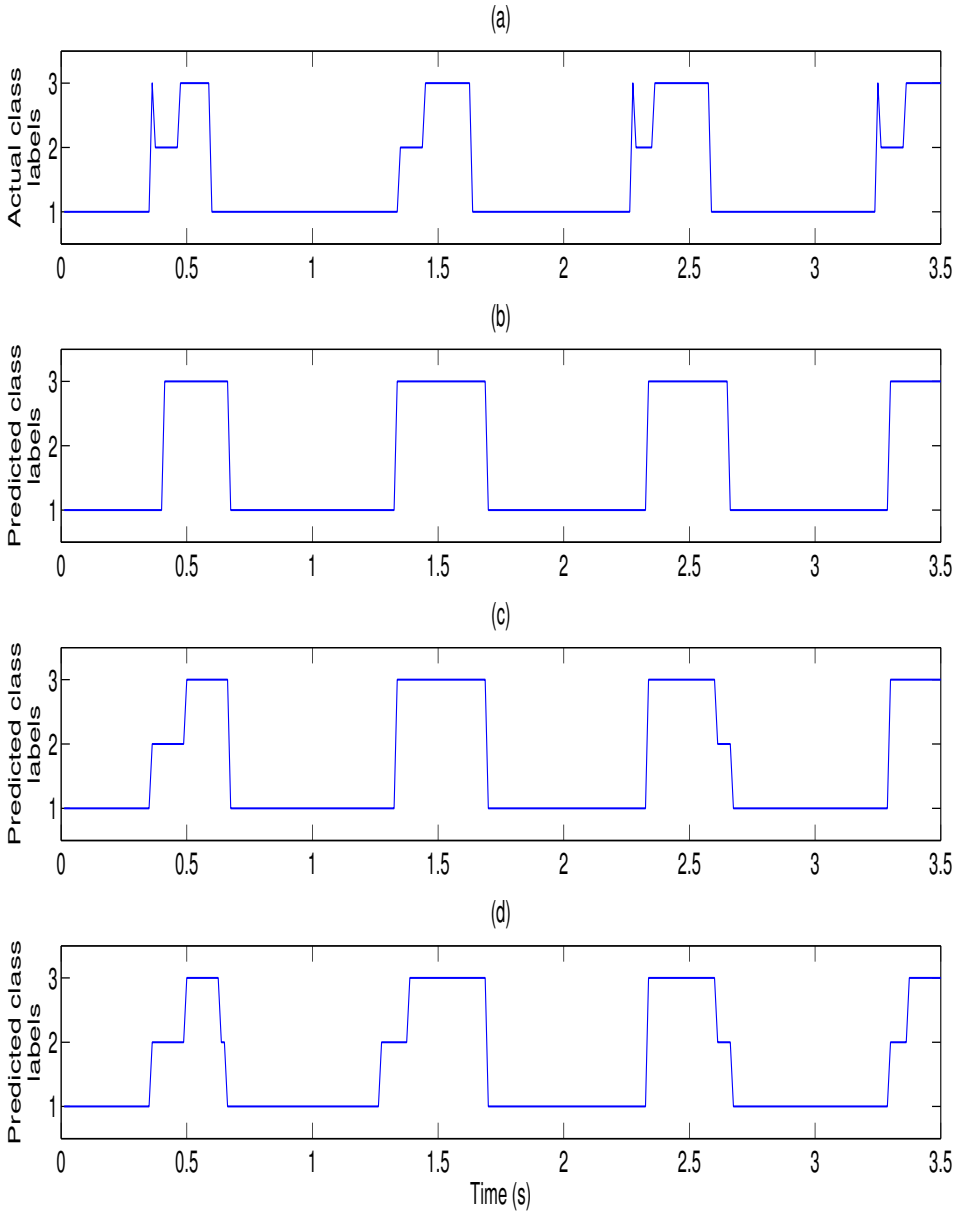

Figure 3.6: Segmentation results for different values of *inter model transition penalty*. a) Actual class labels, b) Predicted class labels for *inter model transition penalty* values -50, c) -20, d) -1. Note that inspiratory phases in the signal are best captured in (d).

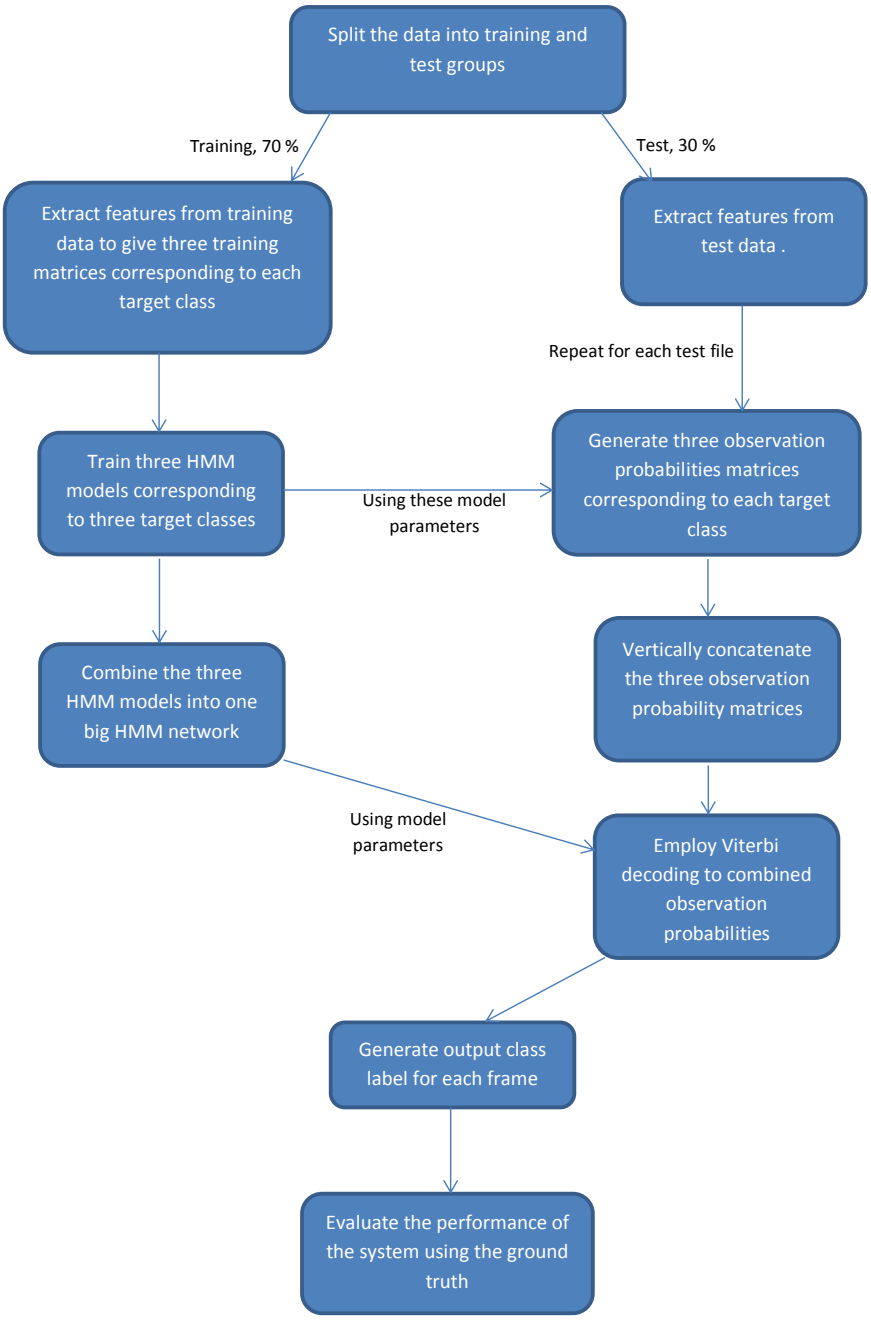

Figure 3.7: Overall audio segmentation system developed for cry signals.

3.2 Fundamental Frequency Estimation

The output of audio segmentation stage for each cry recording under investigation is a sequence of class labels. Based on these labels, desired regions can be identified in the cry recording and acoustic parameters of interest can be extracted. In this thesis, these desired acoustic regions are expiratory and inspiratory phases. Fundamental frequency ($F0$) is an important acoustic parameter whose variation has been found to be correlating with cases of neurological insults in previous infant cry studies [1]. In this thesis we have used YIN algorithm, explained in Section 2.4.4, for $F0$ estimation. A MATLAB implementation of the algorithm freely available at [58] has been used. YIN algorithm has been found to perform well in the context of speech [59] and music signals [60, 61]. But it has not been used in the context of infant cry signals previously.

The output of the algorithm gives a fundamental frequency ($F0$) estimate for each frame of the signal. Here it is assumed that each frame of the cry signal will give a reliable $F0$ which is not true. The $F0$ estimate obtained for a signal frame may or may not be useful depending upon whether or not the signal frame exhibits periodicity. Let us illustrate it by giving example of two signal frames. Figure 3.8 shows an example of magnitude spectrum of a signal frame clearly exhibiting a fundamental frequency $F0$ at 460 Hz and its harmonics at regular intervals equal to $F0$. This $F0$ estimate is certainly reliable.

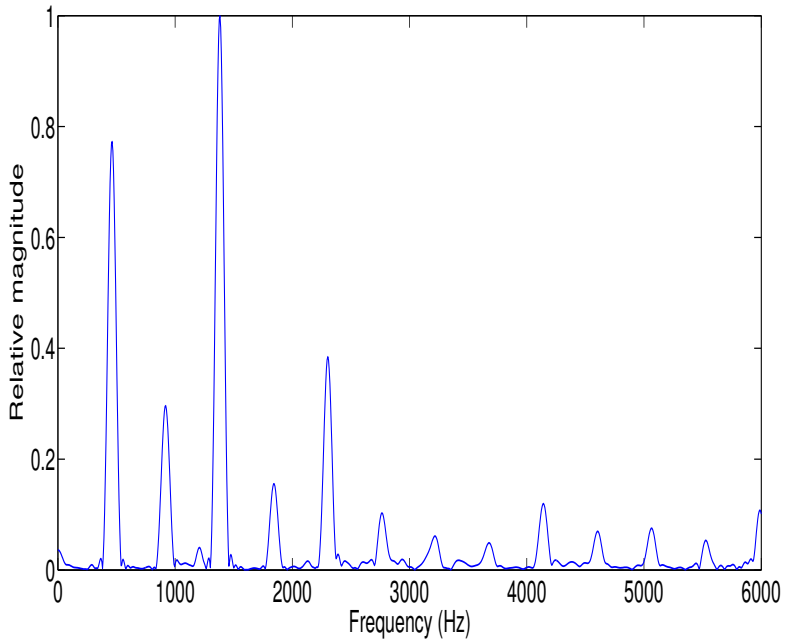

Figure 3.8: Magnitude spectrum of a short time frame of a cry signal. The regular structure of the spectrum exhibits the harmonicity of the signal frame.

Similarly, Figure 3.9 shows an example of magnitude spectrum of an inharmonic frame. It is quite clear that the notion of periodicity for such a frame is not meaningful.

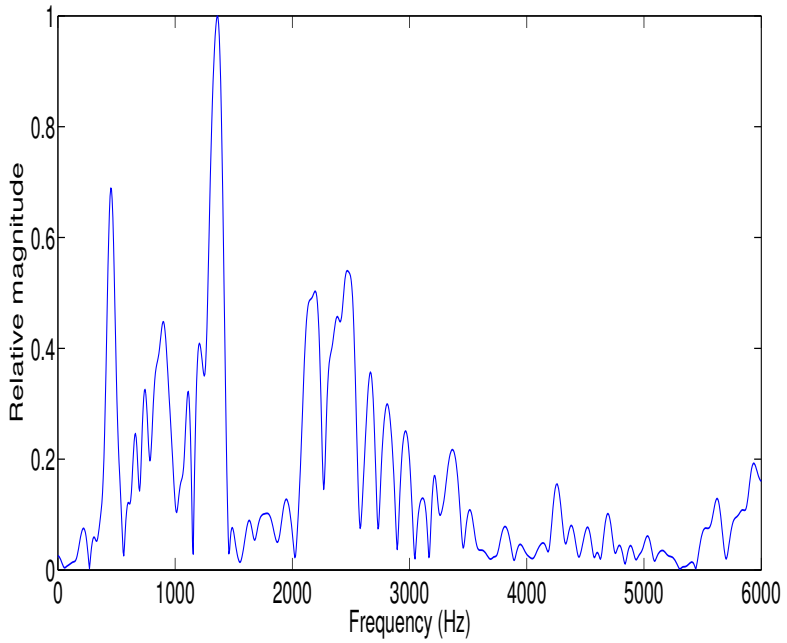

Figure 3.9: Magnitude spectrum of a short time frame of a cry signal. $F0$ estimate for such an irregular spectrum is not meaningful.

Figures 3.8 and 3.9 make it quite clear that there is a need to distinguish the meaningful $F0$ estimates for harmonic frames from the unreliable $F0$ estimates for inharmonic frames.

3.2.1 Refining $F0$ Estimation: Aperiodicity Criterion

In Section 2.4.4, the role of *absolute threshold* on cumulative mean normalized difference function $d'_t(\tau)$, given by Equation (2.25) in restricting the number of possible $T0$ candidates for a particular frame of the signal was explained. It essentially allows for a mechanism against choosing an erroneous high order lag and correspondingly an erroneous low $F0$. This *absolute threshold* can be thought of as controlling the proportion of aperiodic power tolerated in the signal. At time lag equal to time period of the frame $T0$, $d'_t(\tau)$ is proportional to ratio of aperiodic power to total power in the frame [12]. This ratio can be used to distinguish between the two types of $F0$ estimates explained above. We will call this ratio aperiodicity of the frame. The inharmonicity of the frame increases as the value of this ratio increases.

The process of refining the $F0$ estimates, based on aperiodicity, can be summarized in the following two steps:

1. The *absolute threshold* value is used to restrict the list of possible $T0$ candidates for a particular frame. It effectively puts a bar on the proportion of aperiodic power allowed in the frame. Only the $T0$ candidates which fulfill this condition would be eligible for selection. Figure 3.10 shows the output $F0$ values obtained along with the corresponding aperiodicity values for an expiratory phase from a cry recording.

2. In cases where there is no eligible $T0$ candidate, the YIN algorithm outputs the global minimum of $d'_t(\tau)$ as $T0$ estimate. This estimate violates our limit on aperiodicity set via *absolute threshold* parameter and is discarded. After dropping such $T0$ estimates, the output $F0$ values are depicted in Figure 3.10.

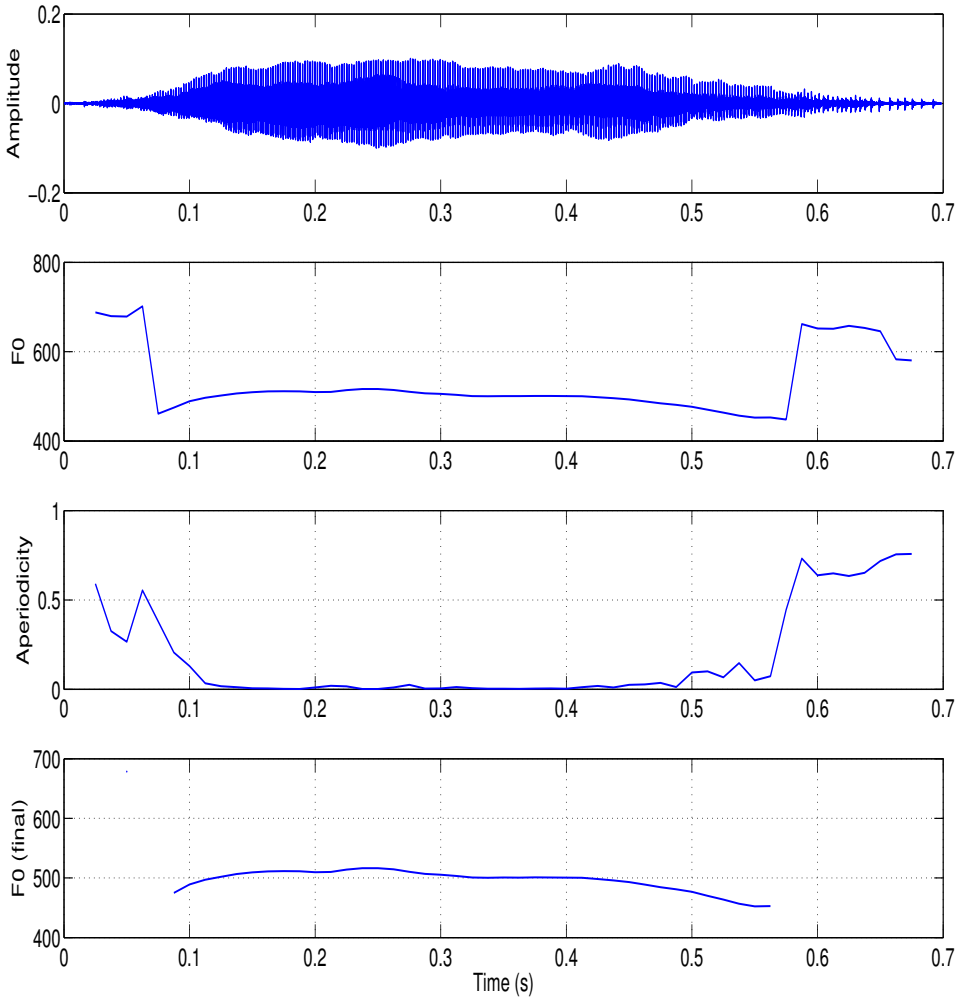

Figure 3.10: Refining $F0$ estimates for an expiratory phase of a cry signal. The top panel shows a chunk of a cry signal under investigation, the second and third panels show the variation of $F0$ and aperiodicity , respectively, obtained for the chunk via YIN algorithm. The bottom panel shows $F0$ values obtained after discarding the frames based on aperiodicity criterion. The *absolute threshold* used is 0.3.

3.2.2 YIN Algorithm Implementation

The overall procedure of $F0$ estimation can thus be listed out in the following steps,

1. YIN algorithm is applied to obtain an $F0$ estimate for each short time frame of a cry signal under investigation. The application of YIN algorithm entails deciding upon several parameters. These are,

    - *Maximum and minimum* $F0$: Setting these parameters restricts the range

for search of the minimum of cumulative mean normalized difference function $d'_t(\tau)$ which gives the $F0$ estimate. In this thesis, we have used the minimum and maximum $F0$ to be 200 and 800 Hz, respectively. Motivation for this are the $F0$ estimation results obtained from previous studies [5, 62].

- *Absolute threshold*: This is a crucial parameter because it is being used here not only for calculating the $F0$ estimates, but also for dropping $F0$ values in inharmonic frames in Step 3. Values from 0.1 to 0.4 have been suggested as good by Cheveigné et al. [12]. Figure 3.11 shows magnitude spectra of four frames of an expiratory phase with aperiodicity values 0.14, 0.28, 0.38, and 0.48, respectively. We have chosen 0.3 as the absolute threshold in this thesis. The motivation for this is making a trade off between dropping too many frames (small aperiodicity/threshold value) versus allowing too many unreliable $F0$ estimates in our analysis (large aperiodicity/threshold value).

2. The class labels obtained in audio segmentation step are used to identify F0 values for expiratory and inspiratory phases.

3. Using the *absolute threshold* decided above, unreliable $F0$ values are discarded.

YIN algorithm is applied for each cry recording and $F0$ estimates are extracted. Their mean and standard deviation can be calculated and are two potential parameters which could be helpful in further analysis involving correlating them with cognitive developmental outcomes. The evolution of $F0$ with time gives rise to $F0$ contours which some researchers have termed as melody of a cry signal [7, 62]. Melody of a cry signal is another potentially useful parameter which would be investigated in this study.

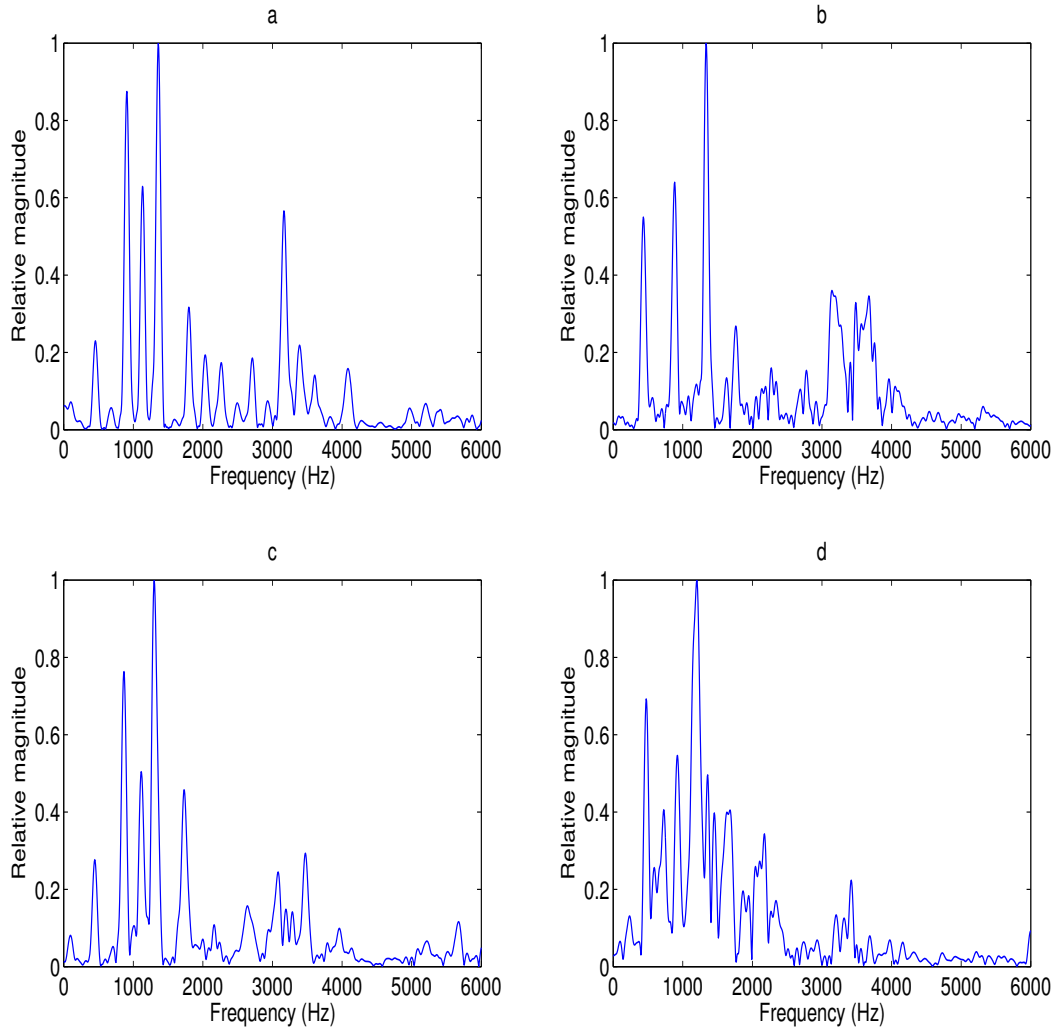

Figure 3.11: Magnitude spectra of four signal frames of an expiratory phase of a cry signal for aperiodicity values, a) 0.14 b) 0.28 c) 0.38 d) 0.48. Note that the inharmonicity of the frame increases with aperiodicity value.

4. EVALUATION

In this chapter, we evaluate the proposed audio segmentation method and report the results. We also report the results for the proposed fundamental frequency estimation method. In the first part, the procedure of collecting data set of cry signal recordings used in this thesis is described. It is followed by a general description of the data set used in audio segmentation, and statistics associated with the distribution of time durations of components of a cry signal found in this data set. Then we explain the performance metrics used to evaluate the audio segmentation results, and performance of the system while using different combinations of audio features and different system configurations. Finally, results of fundamental frequency estimation are discussed, firstly for test data set (chosen for audio segmentation) and then for entire available data set. In the former case, a comparison between the distribution of $F0$ estimate derived on the basis of class information provided by manual annotations is made with the one derived based on class information provided by audio segmentation system.

4.1 Data Collection

The cry recordings for this project were collected at Tampere University Hospital. The recordings were captured by the researchers from Infant Cognition Laboratory, School of Medicine, University of Tampere. A total of 117 infants were included in the study, out of which recordings were successfully captured for 98 infants. The rest either withdrawed or the quality of recording was not good enough to be used in further analysis. Table 4.1 gives the chronological ages of the infants at the time cry recordings were captured.

Table 4.1: The chronological ages of infant subjects

No. of Infants	Chronological age
80	0-3 days
13	4-8 days
5	26-11 weeks[1]

[1]These were older babies that were delivered prematurely

The chronological age of an infant is defined as the time elapsed since birth of the infant. Similarly, Table 4.2 presents the gestational ages of the infants. The gestational age of an infant is defined as the time elapsed between the first day of last menstrual period of the mother and the day of delivery. It is divided into four categories as depicted in Table 4.2.

Table 4.2: The gestational ages of infant subjects.

Gestational age	Time duration	No. of Infants
Very preterm	28-32 weeks	4
Moderate to late preterm	32-37 weeks	4
Term	37-43 weeks	89
Overtime	more than 43 weeks	1

The research study followed the stipulated ethical guidelines and was approved by the Ethical Committee of Tampere University. Consent forms were signed by the guardians of the infants prior to their participation in the study. The recordings were captured in normal hospital environment instead of controlled conditions. The objective was to obtain results which could be useful for developing any possible clinical application in future.

All recordings are 48 kHz sampling rate, two channel audio in 24 bit Waveform Audio file (WAV) format. The audio recorder used was Tascam DR-100MK II with Rode M3 cardiod microphone. The distance between infant's mouth and the recorder was kept at approximately 30 cm . Recordings conditions involved various contexts, e.g., infant crying out of hunger, body temperature being measured, diapers being changed, etc. It sometimes involved pain stimuli like removing a cannula or electrocardiogram (ECG) electrodes from the infant body, or applying venipuncture to draw blood from infant. Each recording is given a separate number code.

4.2 Evaluation: Audio Segmentation

4.2.1 Database

The data set used for audio segmentation consists of 57 cry signal recordings selected from the 98 available recordings. The recordings were selected on the basis of number codes allotted to the cry recordings which correspond to the chronological order in which they were recorded. The first 57 cry codes were chosen and were manually annotated using *Audacity* [?] application. The annotations were done

by the author of this thesis by carefully listening to every cry recording and using subjective judgment to assign class labels. Inspiratory and expiratory phases were quite straightforward to annotate because these sound events are accompanied by characteristic sound of expiration and inspiration produced by an infant. The difficult part was to distinguish non cry vocals from expiratory phases. Expiratory and inspiratory phases generally occur in continuous succession in a cry bout and a cry recording may have several such cry bouts. Non cry vocals generally precede or succeed such cry bouts and are low in amplitude as compared to expiratory phases. This information was used to distinguish them from expiratory phases. In addition, any vocals which did not sound like crying including voices of people talking in the background and other noisy sounds were included in residual class.

Let us look at the three target classes we have in this study: expiratory phases, inspiratory phases and residual class. Expiratory phases consists of expiratory cry sounds with phonation. Expiratory phases without phonation were not found in the data set. Inspiratory phases include inspiratory sounds both with and without phonation. There were very few instances of inspiratory portions without phonation and hence a separate class/label was not created for it. Residual class consists of pauses of no audio activity in between expiratory/inspiratory phases, non cry vocals produced by the infant and other background sounds

The database of 57 manually annotated audio recordings spans around 115 minutes in duration. A total of 1529 expiratory phases were found with mean duration 0.95 s and standard deviation 0.65 s. Figure 4.1 shows the distribution of the time durations for expiratory phases.

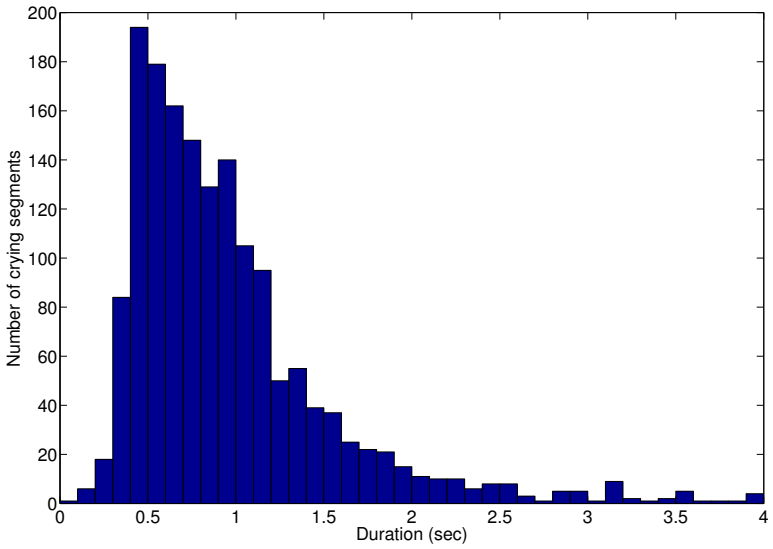

Figure 4.1: The distribution of time durations of expiratory phases.

Similarly, 1005 inspiratory phases were found with mean duration 0.17 s and standard deviation 0.06 s. Figure 4.2 shows the distribution of time durations for inspiratory phases.

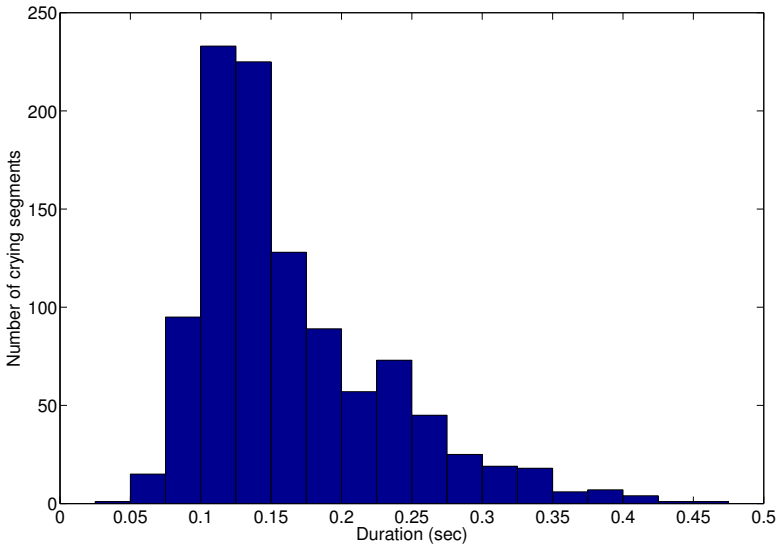

Figure 4.2: The distribution of time durations of inspiratory phases.

The distribution of time durations for expiratory and inspiratory phases is listed out in Table 4.3. As can be inferred from Table 4.3, the inspiratory phases are fewer in number and shorter in duration as compared to the expiratory phases. Hence the data (number of frames) available for training an HMM for inspiratory phases is also lesser as compared to expiratory phases. Moreover, it needs to be emphasized that inspiratory phases also exhibit more variations throughout the data in comparison to expiratory phases. For example, on the one hand, we have audio recordings having very short or almost no discernible inspiratory phases, and on the other hand, we have audio 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 audio recording. Figure 4.3 depicts a portion an audio recording where inspiratory phases is almost absent in comparison to expiratory

Table 4.3: Statistics associated with the distribution of time durations of expiratory and inspiratory phases

Class	No. of segments	Mean duration	Std. deviation	Median
Expiratory phases	1598	0.95 s	0.61 s	0.81 s
Inspiratory phases	1042	0.16 s	0.07 s	0.14 s

phases, and Figure 4.4 depicts the opposite case where these are quite prominent.

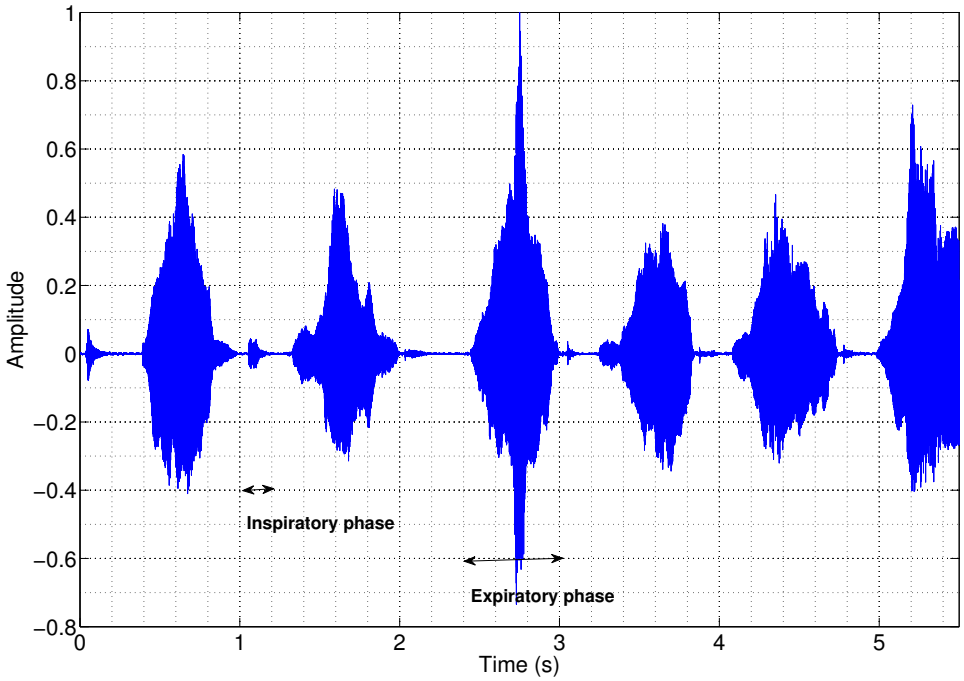

Figure 4.3: An example of a chunk of a cry signal with inconspicuous inspiratory phases.

This wide variation in inspiratory phases is quite challenging to deal with while training the corresponding HMM for audio segmentation. We have too few data available for training and the data exhibits a wide range of variation across the available data set. This observation is reflected in the poor performance of the audio segmentation system for inspiratory phases as compared to expiratory phases. Section 4.2.3 discusses the performance of the system for both these classes with different configurations of the HMM states and number of component Gaussians.

As explained in Section 3.1, the available data set of 57 cry recordings is split into training and test sets. First 70% of the available annotated data, i.e., 40 audio files were selected for training the HMMs and the remaining 30% of the available data, i.e., 17 files were used for testing the model. As the cry recordings are numbered according to chronological order, the training data consists of files captured earlier than test data recordings.

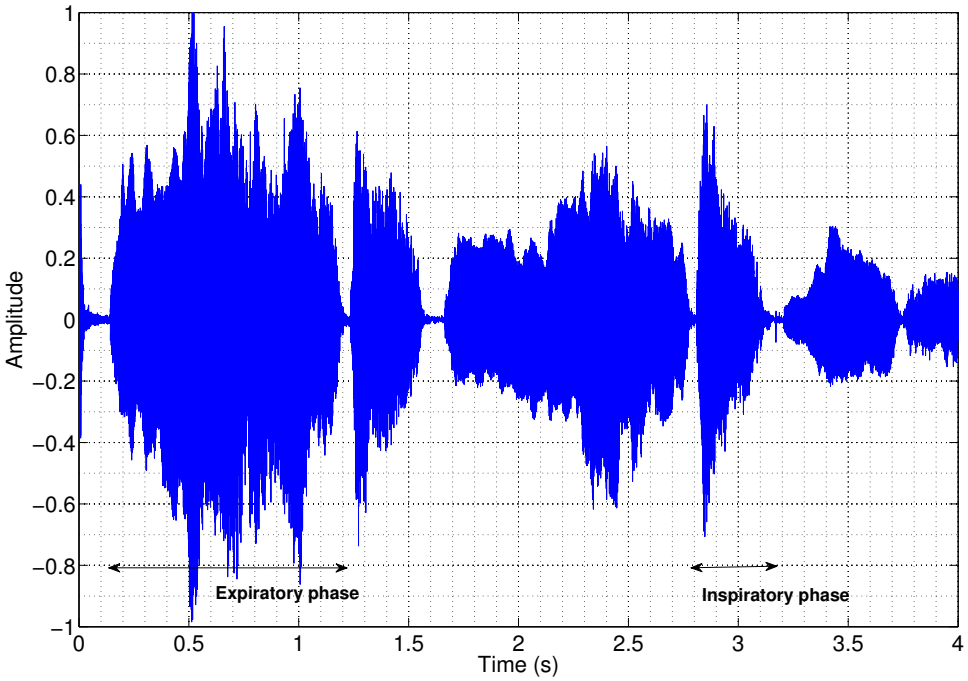

Figure 4.4: An example of a chunk of a cry signal with prominent inspiratory phases.

4.2.2 Performance Metrics

The HMM pattern recognizer is evaluated on a test set and the output labels produced by it are compared against the ground truth. The ground truth in this thesis are the manual annotations obtained via *Audacity* [**?**] application as described in Chapter 3. There are two metrics which have been used in this thesis to evaluate the performance of the system, namely, accuracy and F score. Note that both of these metrics are frame based in this thesis.

1. *Accuracy*: The frame based accuracy is defined as the ratio of correctly labeled frames to the total number of frames in a signal. A correctly labeled frame is one for which output label generated by the pattern recognizer matches with the true label learned from the ground truth. For binary classification problem, accuracy can be defined for each target class, but for multi-class classification problem it can only be defined for the overall system. Accuracy can be calculated using,

$$accuracy = \frac{number\ of\ correctly\ classified\ frames}{total\ number\ of\ frames} \quad . \tag{4.1}$$

2. *F score*: The F score is defined as the harmonic mean of precision and recall values. Precision, also known as positive predictive value, 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.

   Precision can be calculated using,

$$P = \frac{number\ of\ true\ positive\ frames}{number\ of\ test\ outcome\ positive\ frames} \quad . \tag{4.2}$$

   Recall, also known as true positive rate or sensitivity of the system, 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. Hence, it is the number of actual positive frames for a particular class. Recall can be calculated, using

$$R = \frac{number\ of\ true\ positive\ frames}{number\ of\ actual\ positive\ frames} \quad . \tag{4.3}$$

   Note that both Precision and Recall are defined with respect to a particular class. The same stands true for the F score measure as well. Using Equations (4.2) and (4.3), F score can be calculated as,

$$F = 2\,\frac{P \cdot R}{P + R} \quad . \tag{4.4}$$

These performance metrics are calculated for all the available test files. The final performance metric for the system is given by the average of metrics calculated for the individual files.

4.2.3 Varying HMM States and Number of Gaussians

In Chapter 3, it was discussed that the HMM pattern recognizer system consisting of three HMM models corresponding to three target classes is trained with different number of states and component Gaussians. In this section, we will first describe the baseline configuration of states and component Gaussians. The performance of this baseline configuration is then compared with more sophisticated configurations involving more HMM states and number of Gaussians. Features used in this baseline configuration are MFCCs.

In the baseline configuration, each HMM model corresponding to one of the three target classes is trained with 1 state and 5 Gaussians. We observed that increasing the number of Gaussians improves the overall accuracy of the classification. Corresponding improvements in the F scores of expiratory and inspiratory phases was also observed. The improvement in performance metrics while going from 15 component Gaussians to 20 component Gaussians was however barely noticeable. Table 4.4 gives the system accuracy and F scores with varying number of component Gaussians. An accuracy of 85.3 % and corresponding F scores of 82.37 % for expiratory phases and 38.6 % for inspiratory phases were obtained for this system configuration consisting of 1 HMM state and 15 Gaussian components for each class. HMM configurations with one state and varying number of Gaussian components has been previously explored in [27], where best average accuracy of 86.4 % has been reported for 20 Gaussian components using MFCC features.

Table 4.4: The performance of the system with different number of component Gaussians

Features	No. of Gaussians per state	Accuracy (%)	F score (%)	
			Inspiratory phases	Expiratory phases
MFCCs	5	84.2	37.6	81.4
MFCCs	10	84.7	38.5	82.3
MFCCs	15	85.3	38.6	82.7
MFCCs	20	85.3	38.7	82.6

Similarly, the effect of using more than one HMM state for the three target classes was investigated. Different number of HMM states and number of component Gaussians were experimented with. The accuracies and F scores of the model are reported in Table 4.5. It can be observed that using more than one HMM states the system accuracies can be improved up to 87.5 %. The best performance is achieved for 2, 1, and 3 HMM states corresponding to expiratory phase, inspiratory phase and residual classes, respectively, with each of them composed of 10 component Gaussians.

Table 4.5: The performance of the system with different number of HMM states and component Gaussians using MFCC features

No. of States			No. of Gaussians			F score (%)		Accuracy
Exp[2]	Insp[3]	Res[4]	Exp[2]	Insp[3]	Res[4]	Insp[3]	Exp[2]	(%)
2	1	1	5	5	20	37.1	81.9	84.6
3	1	3	4	4	4	41.2	83.7	86.6
2	2	2	10	10	10	41.2	83.2	85.8
2	1	1	10	20	10	39.5	81.5	84.8
2	1	1	10	20	20	39.7	82.3	85.5
2	1	2	10	5	10	42.6	83.6	87.0
2	1	2	10	10	10	42.9	83.6	87.1
2	**1**	**3**	**10**	**10**	**10**	**44.0**	**83.7**	**87.5**
3	1	2	10	20	10	42.4	82.2	86.3
3	1	3	10	10	10	42.1	83.2	86.9

4.2.4 Use of Additional Features

In this section, we will describe the system performance with use of additional audio features along with MFCCs. The best known configuration of HMM states and component Gaussians learned from the previous section was experimented with using different combination of audio features. The reference system configuration thus consists of 2,1, and 3 HMM states for expiratory phase, inspiratory phase and residual class, respectively, and each of the target class consists of 10 component Gaussians. The following features were experimented with,

1. *Delta coefficients and delta-delta coefficients*: A combination of delta and delta-delta features with MFCCs resulted in improved F score performance of the system for inspiratory phase class. We were able to achieve 50 % F scores with delta-delta features. Table 4.6 describes the system's F score performances and obtained accuracies. The overall accuracy of the system was also improved to 88 %.

[2]Expiratory phases
[3]Inspiratory phases
[4]Residual

2. *Running average and running variance of MFCCs*: The running averages and running variances of MFCC features were calculated for sliding windows of size 5, 10 and 15 frames. The accuracies obtained were poorer as compared to the ones achieved MFCCs alone. However, improvements were observed in F score performance for inspiratory phase class. This improvement, however is poor compared to one achieved with delta and delta-delta coefficients. Table 4.6 describes the F score performances and obtained accuracies for the system. The window length for calculating running averages and running variances are indicated as well.

3. *Fundamental frequency of the signal in each frame*: An improvement in the F score performance of the inspiratory phase class was observed by including fundamental frequency of the frames as feature with MFCCs. A slight improvement in overall accuracy of the system is observed as well.

4. *Aperiodicity of the signal in each frame*: An improvement in the F score performance of both expiratory and inspiratory phases were observed by including aperiodicity of the frames as feature with MFCCs. The overall accuracy of the system was also improved to 88.0 %.

As is evident from the Table 4.6, use of deltas, delta-deltas; $F0$; and aperiodicity features along with MFCCs led to 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 combination of MFCCs with deltas and delta-deltas yielded most improvement in F score performance of inspiratory phases. Hence, this combination is further investigated with $F0$ and aperiodicity audio features. The obtained results are reported in Table 4.6. The overall accuracy of the system was improved up to 88.5 % for a combination of MFCCs; deltas, delta-deltas; and aperiodicity, with a corresponding improvement in the F score performance, namely, 52.0 % for inspiratory phases and 84.8 % for expiratory phases.

Table 4.6: The performance of the model with additional features

Features	Acccuracy (%)	F score (%)	
		Insp[3]	Exp[2]
MFCCs	87.5	44.0	83.7
MFCCs + deltas and delta-deltas	88.0	50.5	84.3
MFCCs + running averages and running variances (5 frames)	86.6	43.5	84.3
MFCCs + running averages and running variances (10 frames)	86.1	44.3	83.0
MFCCs + running averages and running variances (15 frames)	85.6	42.5	82.5
MFCCs + $F0$	88.1	51.3	83.8
MFCCs + $F0$ + deltas and delta-deltas	88.2	50.5	84.2
MFCCs + aperiodicity	88.0	49.8	85.1
MFCCs + aperiodicity +deltas and delta-deltas	**88.5**	**52.0**	**84.8**

4.3 Results: $F0$ Estimation

In this section, results obtained from fundamental frequency estimation of the cry recordings will be presented. $F0$ estimates for each cry recording were collected using YIN algorithm and post-processed via the aperiodicity criterion as explained in Section 3.2.2. $F0$ estimate statistics, i.e., mean, median and standard deviations are computed and reported here.

4.3.1 $F0$ Estimation for Test Data set

We have a database of 98 infants out of which 57 have been manually annotated for class labels. Moreover, in the audio segmentation step, we divided this manually annotated data into training and test data sets consisting of 40 and 17 recordings,

respectively. In this section, we will describe the distribution of $F0$ estimates obtained for the test data set. The objective is to compare the $F0$ estimates derived on the basis of manual annotations with the ones derived on the basis of audio segmentation results. Figures 4.5 and 4.6 show the distribution of $F0$ estimates derived on the basis of class labels provided by manual annotations for expiratory and inspiratory phases, respectively.

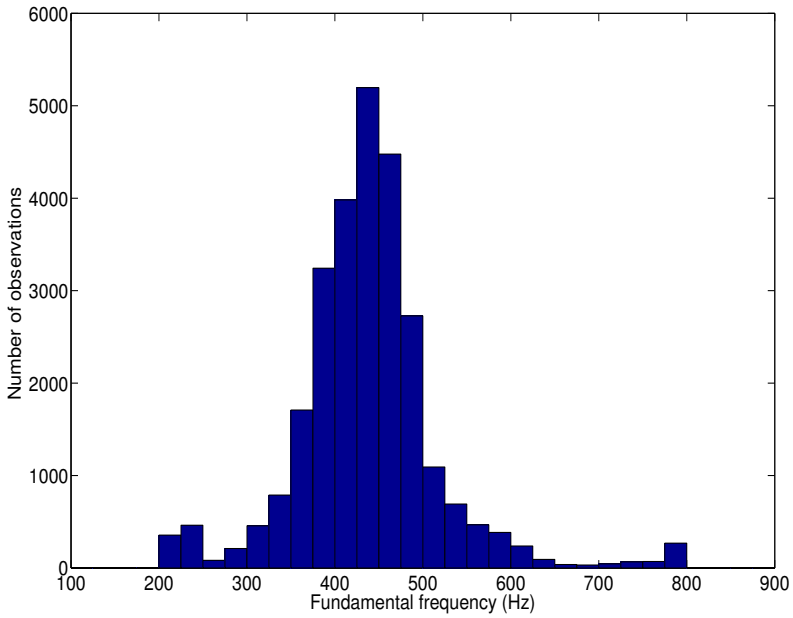

Figure 4.5: The distribution of $F0$ estimates for expiratory phases derived from the test data set. The class information used for their extraction is provided by manual annotations.

Table 4.7 shows the statistics derived from distributions shown in Figures 4.5 and 4.6. Note that mean, standard deviation, and median values are reported for $F0$ estimates collected for the entire test data set while maximum and minimum mean values are reported for individual cry recordings.

Table 4.7: *F0* statistics derived from test data on the basis of manually annotated classes (in Hz)

Class	Mean	Std. dev.	Median	Max. mean	Min. mean
Expiratory phases	437.4	81.8	437.2	503.5	391.1
Inspiratory phases	579.9	148.4	579.9	764.6	442.7

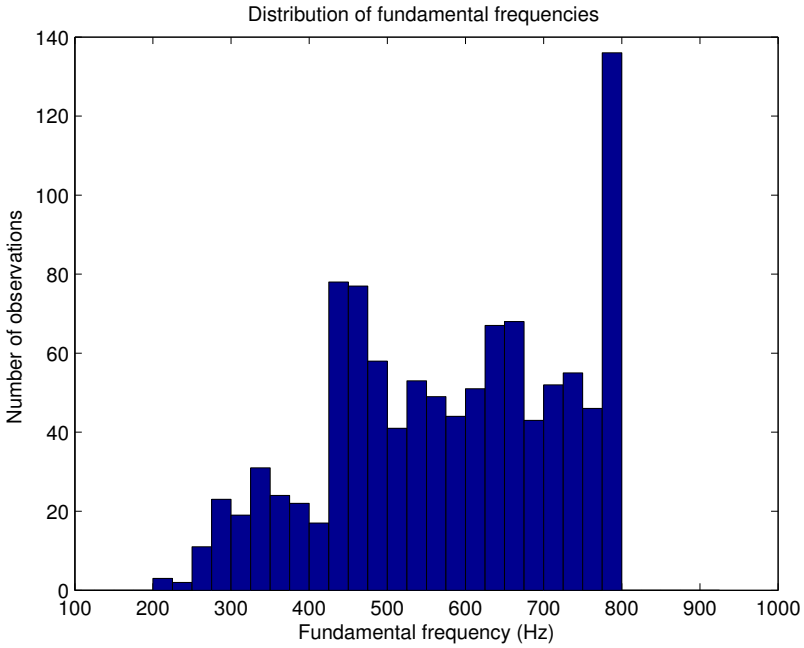

Figure 4.6: The distribution of $F0$ estimates for inspiratory phases derived from the test data set. The class information used for their extraction is provided by manual annotations.

Similarly, Figures 4.7 and 4.8 depict the distributions of $F0$ estimates derived on the basis of class information provided by audio segmentation system for expiratory and inspiratory phases, respectively. The associated statistics are given in Table 4.8. Here also the mean, standard deviation, and median values are reported for the $F0$ estimates collected for the entire test data-set while maximum and minimum mean values are reported for individual cry recordings.

Table 4.8: *F0* statistics derived from test data based on segmentation results (in Hz)

Class	Mean	Std. dev.	Median	Max. mean	Min. mean
Expiratory phases	441.9	84.4	439.2	506.8	392.3
Inspiratory phases	499.5	145.2	451	623.6	303.7

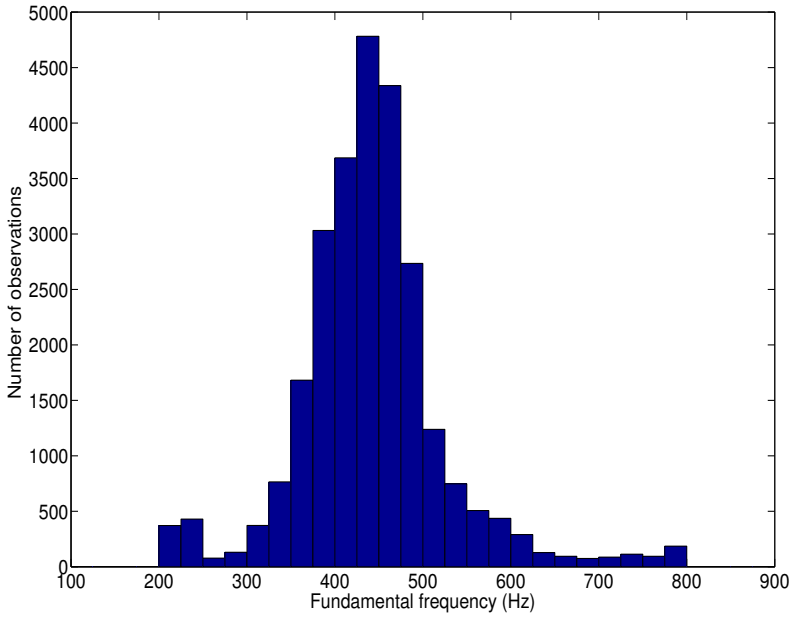

Figure 4.7: The distribution of $F0$ estimates for expiratory phases derived from the test data set. The class information used for their extraction is provided by audio segmentation results.

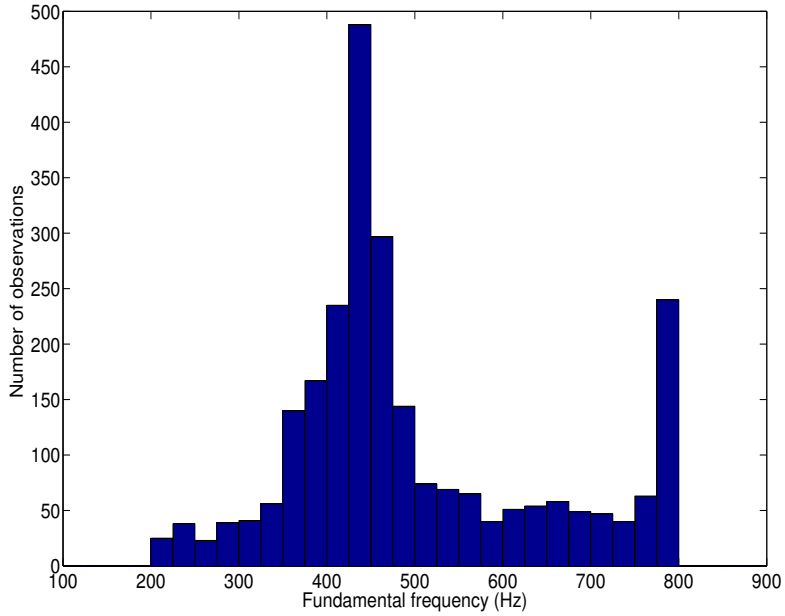

Figure 4.8: The distribution of $F0$ estimates for inspiratory phases derived from the test data set. The class information used for their extraction is provided by audio segmentation results.

Tables 4.7 and 4.8 show that for expiratory phases the derived $F0$ estimate statistics are quite similar. Although, the same is not true for inspiratory phases. This can also be observed by comparing $F0$ distributions for expiratory phases (Figures 4.5 and 4.7) and inspiratory phases (Figures 4.6 and 4.8). In fact, the distribution of $F0$ estimates for inspiratory phases resembles that of expiratory phases due to a lot of false positives from the former distribution leaking into the latter. This further underlines the earlier observation that the audio segmentation works well with expiratory phases but performs poorly for inspiratory phases. The diverse nature of inspiratory phases present in our data set, explained in Section 4.2.1, is responsible for this poor performance.

4.3.2 $F0$ Estimation for Entire Data set

In this section, the results obtained from $F0$ estimation of the entire available data set of cry recordings i.e., 98 recordings will be presented. The estimates are derived based on the class information provided by audio segmentation system. Figures 4.9 and 4.10 show the distribution of $F0$ estimates for expiratory and inspiratory phases, respectively. The associated statistics are given by Table 4.9. We can again see the distribution of inspiratory phases resembling that of expiratory phases due to false positives in the audio segmentation outputs.

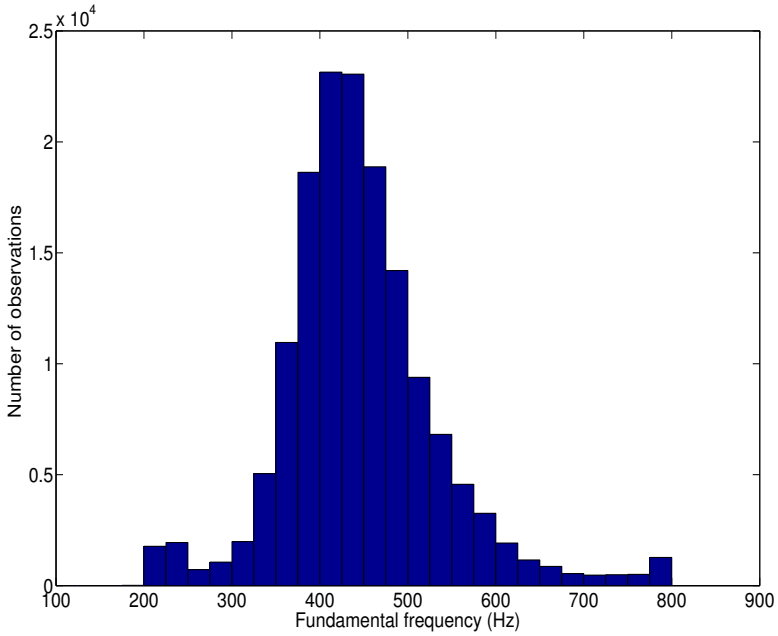

Figure 4.9: The distribution of $F0$ estimates for expiratory phases derived from the entire available data set. The class information used for their extraction is provided by audio segmentation results.

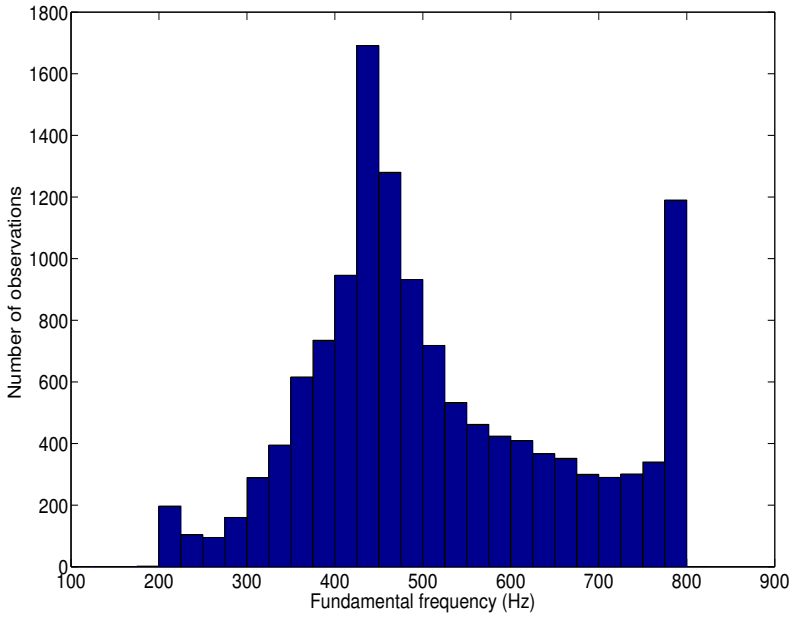

Figure 4.10: The distribution of $F0$ estimates for inspiratory phases derived from the entire available data set. The class information used for their extraction is provided by audio segmentation results.

Table 4.9: *F0* statistics derived from entire available data set based on class information derived from audio segmentation results (in Hz)

Class	Mean	Std. dev.	Median	Max. mean	Min. mean
Expiratory phases	445.7	88.2	437.2	634.5	331.3
Inspiratory phases	514.8	146.0	476.5	677.3	272.8

We can alternatively calculate the mean $F0$ estimate for each cry recording in order to reveal the $F0$ characteristics associated with each infant. Figures 4.11 and 4.12 show the distribution of file based $F0$ means for expiratory and inspiratory phases, respectively.

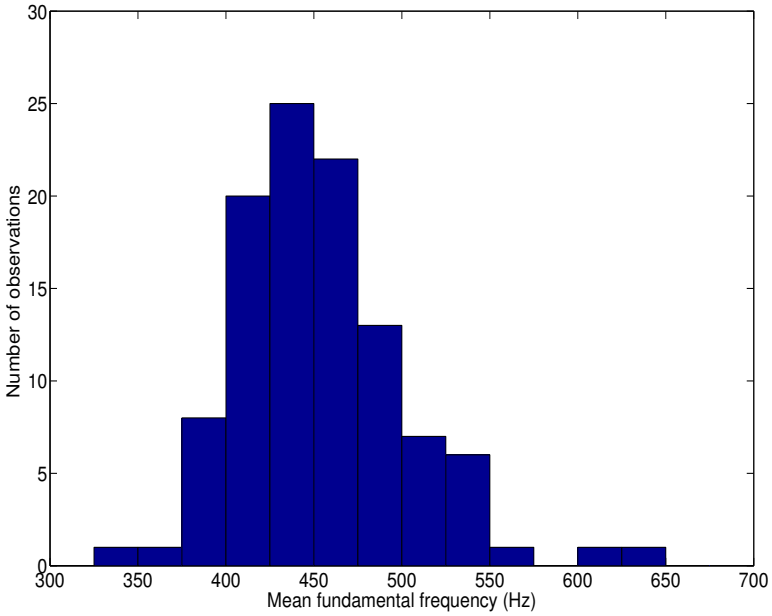

Figure 4.11: The distribution of mean $F0$ estimates for expiratory phases derived for individual cry recordings. The class information used for their extraction is provided by audio segmentation results.The mean of this distribution is 449.3 Hz.

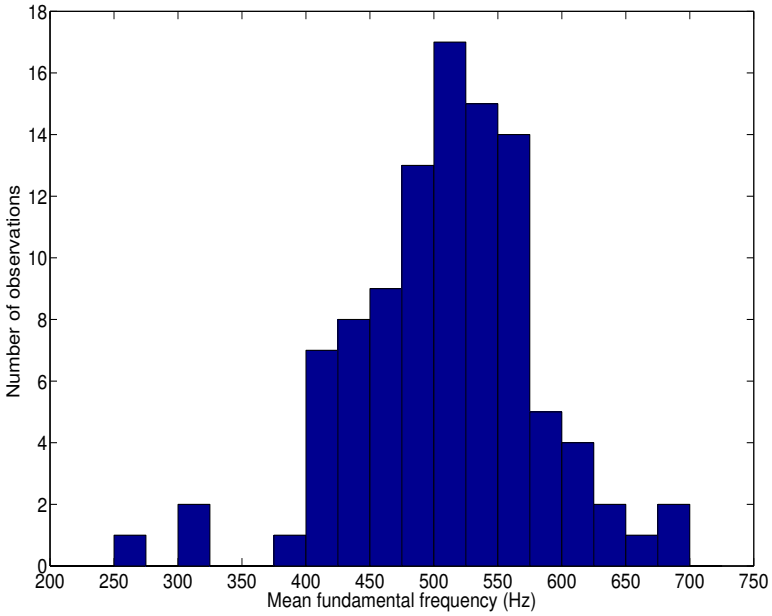

Figure 4.12: The distribution of mean $F0$ estimates for inspiratory phases derived for individual cry recordings. The class information used for their extraction is provided by audio segmentation results. The mean of this distribution is 505.3 Hz.

In a similar fashion, file based variances and standard deviations of $F0$ estimates can be calculated to reveal $F0$ variation within a file. We can even go further and investigate micro level behavior of $F0$ estimates on the level of individual expiratory and inspiratory phases. These file based $F0$ statistics along with individual expiratory/inspiratory phase level $F0$ behavior within each file would be most useful in investigating the correlations with cognitive developmental outcomes of the infants.

5. CONCLUSIONS AND FUTURE WORK

This thesis is a part of an ongoing study on infant cry analysis funded by Centre for Child Health Research, University of Tampere. It is being contributed to by Audio Research Group, Tampere University of Technology; and Infant Cognition Laboratory, University of Tampere. This study aims to analyze infant cry recordings in order to find potential markers which would help in early assessment of neurological development problems in infants. The present thesis is focused on two research problems in the context of infant cry signals: audio segmentation of infant cry recordings in order to extract audio parts that are relevant to further analysis and fundamental frequency ($F0$) estimation of these extracted relevant parts. The relevant parts here are expiratory and inspiratory phases. Fundamental frequency ($F0$) is an important acoustic parameter of cry signals which has been found to be useful in previous infant cry studies, and meaningful correlations have been drawn between its behavior and cases of neurological insults.

The experiments have been conducted on an audio database consisting of cry signal recordings captured in a realistic hospital environment. The recording conditions involved varied contexts like hunger cries, pain cries while applying venipuncture or removing ECG electrodes from infant body, cries recorded while measuring infant body temperature, cries recorded while changing diapers, etc. The diverse nature of the contexts incorporated in the database used in this thesis distinguishes this study from similar attempts done previously which have been mostly concentrated on specific contexts.

An HMM based audio segmentation system has been proposed as a solution for the audio segmentation problem which works well for expiratory phases but performs poorly for inspiratory phases. The reason for this is the diverse nature of inspiratory phases present in our data set. Some of the audio recordings exhibit quite prominent inspiratory phases, in some cases even more prominent than the expiratory phases, while others exhibit their complete absence altogether. The chosen performance metrics to describe the system efficiency are frame based accuracy and frame based F scores. Various configurations of HMM states and number of component Gaussians were experimented with. Using MFCC features, the best performing configuration involved 2,3, and 1 HMM states for expiratory phases, inspiratory phases and residual classes, respectively, with each HMM state consisting of 10 component

Gaussians. An accuracy of 87.5 % was achieved for this configuration which corresponds to F scores of 44 % and 83.7 % for expiratory and inspiratory phases, respectively. Various combinations of audio features, namely, MFCCs, deltas, delta-deltas, running averages, running variances, fundamental frequency, and aperiodicity were experimented with. The best performing feature combination employed MFCCs, aperiodicity, deltas and delta-deltas. The final accuracy of 88.5 % was achieved for this configuration and feature combination which corresponds to F scores of 53.3 % and 84.7 % for expiratory and inspiratory phases, respectively.

YIN algorithm is applied in order to solve the fundamental frequency ($F0$) estimation problem. Through this method, an $F0$ estimate is obtained for each frame of a signal irrespective of the harmonicity of the concerned frame. Aperiodicity, which is proportional to the amount of aperiodic power contained in a signal frame, is the measure of inharmonicity of the signal frame. It can be used to discard the unreliable $F0$ estimates. Use of aperiodicity to refine $F0$ estimates has been referred to as the aperiodicity criterion in this thesis. The statistics associated with the distribution of $F0$ estimates corresponding to expiratory and inspiratory phases are reported.

The application of YIN algorithm for $F0$ estimation in the context of cry signals is novel in this thesis. One direction of future research could be the evaluation of this algorithm against the other $F0$ estimation algorithms used previously in infant cry research in order to ascertain how it fares in comparison.

The cognitive developmental outcomes for the infants are being collected at the time this thesis is being written. It has two components: health data and eye tracking data. The health data consists of collection of risk factors associated with the infant as well as risk factors associated with the mother during pregnancy and at delivery. It also consists of diagnosis of health and developmental problems of the infant after delivery and on the day the cry recordings were captured. The eye tracking data measures the oculomotor orientation and attentional focus of the infant when presented with some visual stimuli. It serves as an early marker of cognitive development of the infant. The future work would involve correlating the $F0$ estimates and their variations within a cry recording or may be within an expiratory/inspiratory phase with the cognitive developmental outcomes of the infants. Moreover, other acoustic parameters, e.g., formants, duration of expiratory/ inspiratory phases, amplitudes of expiratory/ inspiratory phases, etc. can be investigated and their correlation with the cognitive developmental outcomes can be studied.

REFERENCES

[1] L. L. LaGasse, A. R. Neal, and B. M. Lester, "Assessment of infant cry: acoustic cry analysis and parental perception," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 11, no. 1, pp. 83–93, 2005.

[2] K. Michelsson and O. Michelsson, "Phonation in the newborn, infant cry," *International Journal of Pediatric Otorhinolaryngology*, vol. 49, pp. 297–301, 1999.

[3] G. Várallyay, A. Illényi, and Z. Benyó, "Automatic infant cry detection.," in *Models and Analysis of Vocal Emissions for Biomedical Applications*, pp. 11–14, 2009.

[4] G. Esposito, M. del Carmen Rostagno, P. Venuti, J. D. Haltigan, and D. S. Messinger, "Brief report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for asd," *Journal of Autism and Developmental Disorders*, vol. 44, no. 4, pp. 975–980, 2014.

[5] H. Rothgänger, "Analysis of the sounds of the child in the first year of age and a comparison to the language," *Early Human Development*, vol. 75, no. 1, pp. 55–69, 2003.

[6] M. J. Corwin, B. M. Lester, C. Sepkoski, M. Peucker, H. Kayne, and H. L. Golub, "Newborn acoustic cry characteristics of infants subsequently dying of sudden infant death syndrome," *Pediatrics*, vol. 96, no. 1, pp. 73–77, 1995.

[7] K. Wermke, M. Birr, C. Voelter, W. Shehata-Dieler, A. Jurkutat, P. Wermke, and A. Stellzig-Eisenhauer, "Cry melody in 2-month-old infants with and without clefts," *The Cleft Palate-Craniofacial Journal*, vol. 48, no. 3, pp. 321–330, 2011.

[8] K. Michelsson, P. Sirviö, M. Koivisto, A. Sovijarvi, and O. Wasz-Hockert, "Spectrographic analysis of pain cry in neonates with cleft palate," *Neonatology*, vol. 26, no. 5-6, pp. 353–358, 1975.

[9] H. L. Golub and M. J. Corwin, "Infant cry: a clue to diagnosis," *Pediatrics*, vol. 69, no. 2, pp. 197–201, 1982.

[10] R. Colton and A. Steinschneider, "The cry characteristics of an infant who died of the sudden infant death syndrome," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, pp. 359–363, 1981.

[11] S. M. Grau, M. P. Robb, and A. T. Cacace, "Acoustic correlates of inspiratory phonation during infant cry," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 2, pp. 373–381, 1995.

[12] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[13] O. Wasz-Höckert, K. Michelsson, and J. Lind, "Twenty-five years of Scandinavian cry research," in *Infant Crying*, pp. 83–104, Springer, 1985.

[14] K. Michelsson, "Cry analyses of symptomless low birth weight neonates and of asphyxiated newborn infants," *Acta Pædiatrica*, vol. 60, no. S216, pp. 9–45, 1971.

[15] P. Sirviö and K. Michelsson, "Sound-spectrographic cry analysis of normal and abnormal newborn infants," *Folia Phoniatrica et Logopaedica*, vol. 28, no. 3, pp. 161–173, 1976.

[16] H. Golub and M. Corwin, "A physioacoustic model of the infant cry," in *Infant Crying* (B. Lester and C. Zachariah Boukydis, eds.), pp. 59–82, Springer US, 1985.

[17] K. Michelsson and P. Sirvio, "Cry analysis in herpes encephalitis," in *Proceedings of the Fifth Scandinavian Congress on Perinatal Medicine*, 1975.

[18] J. Lind, V. Vuorenkoski, G. Rosberg, T. Partanen, and O. Wasz-Hockert, "Spectographic analysis of vocal response to pain stimuli in infants with down's syndrome," *Developmental Medicine & Child Neurology*, vol. 12, no. 4, pp. 478–486, 1970.

[19] V. R. Fisichelli and S. Karelitz, "Frequency spectra of the cries of normal infants and those with down's syndrome," *Psychonomic Science*, vol. 6, no. 4, pp. 195–196, 1966.

[20] V. Vuorenkoski, J. Lind, T. Partanen, J. Lejeune, J. Lafourcade, and O. Wasz-Hockert, "Spectrographic analysis of cries from children with maladie du cri du chat.," in *Annales Paediatriae Fenniae*, vol. 12, pp. 174–180, 1966.

[21] S. Karelitz and V. R. Fisichelli, "The cry thresholds of normal infants and those with brain damage: An aid in the early diagnosis of severe brain damage," *The Journal of P ediatrics*, vol. 61, no. 5, pp. 679–685, 1962.

[22] J. Lind, O. Wasz-Höckert, V. Vuorenkoski, and E. Valanne, "The vocalization of a newborn, brain-damaged child," in *Annales paediatriae Fenniae*, vol. 11, pp. 32–37, 1964.

[23] B. Reggiannini, S. J. Sheinkopf, H. F. Silverman, X. Li, and B. M. Lester, "A flexible analysis tool for the quantitative acoustic assessment of infant cry," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1416–1428, 2013.

[24] C. Manfredi, L. Bocchi, S. Orlandi, L. Spaccaterra, and G. Donzelli, "High-resolution cry analysis in preterm newborn infants," *Medical Engineering & Physics*, vol. 31, no. 5, pp. 528–532, 2009.

[25] A. Messaoud and C. Tadj, "A cry-based babies identification system," in *Image and Signal Processing*, pp. 192–199, Springer, 2010.

[26] Y. Kheddache and C. Tadj, "Acoustic measures of the cry characteristics of healthy newborns and newborns with pathologies," *Journal of Biomedical Science and Engineering*, vol. 2013, 2013.

[27] J.-J. Aucouturier, Y. Nonaka, K. Katahira, and K. Okanoya, "Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden markov models," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2969–2977, 2011.

[28] B. M. Lester and P. S. Zeskind, "A biobehavioral perspective on crying in early infancy," in *Theory and Research in Behavioral Pediatrics*, pp. 133–180, Springer, 1982.

[29] F. L. Porter, R. H. Miller, and R. E. Marshall, "Neonatal pain cries: effect of circumcision on acoustic features and perceived urgency," *Child Development*, pp. 790–802, 1986.

[30] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." `http://www.praat.org`. Online; accessed 2015-02-21.

[31] "Computerized Speech Lab (CSL) - KayPENTAX." `http://kayelemetrics.com`. Online; accessed 2015-02-21.

[32] H. and M. Souza, "Study of acoustic features of newborn cries that correlate with the context," in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, vol. 3, pp. 2174–2177, IEEE, 2001.

[33] H. C. Lin and J. A. Green, "Effects of posture on newborn crying," *Infancy*, vol. 11, no. 2, pp. 175–189, 2007.

[34] J. R. Irwin, "Parent and nonparent perception of the multimodal infant cry," *Infancy*, vol. 4, no. 4, pp. 503–516, 2003.

[35] L. Rautava, A. Lempinen, S. Ojala, R. Parkkola, H. Rikalainen, H. Lapinleimu, L. Haataja, L. Lehtonen, P. S. Group, *et al.*, "Acoustic quality of cry in very-low-birth-weight infants at the age of 1 1/2 years," *Early Human Development*, vol. 83, no. 1, pp. 5–12, 2007.

[36] B. Mampe, A. D. Friederici, A. Christophe, and K. Wermke, "Newborns' cry melody is shaped by their native language," *Current Biology*, vol. 19, no. 23, pp. 1994–1997, 2009.

[37] J. Markel, "The sift algorithm for fundamental frequency estimation," *Audio and Electroacoustics, IEEE Transactions on*, vol. 20, no. 5, pp. 367–377, 1972.

[38] D. Lederman, "Estimation of infants' cry fundamental frequency using a modified sift algorithm," *arXiv preprint arXiv:1009.2796*, 2010.

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

[40] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000.

[41] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[42] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[43] L. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[44] J. C. Segura, A. J. Rubio, A. M. Peinado, P. Garcia, and R. Román, "Multiple VQ hidden Markov modelling for speech recognition," *Speech Communication*, vol. 14, no. 2, pp. 163–170, 1994.

[45] K. Lee, *Automatic Speech Recognition: The Development of the SPHINX Recognition System*. Kluwer international series in engineering and computer science: VLSI, computer architecture, and digital signal processing, Springer, 1989.

[46] X. Huang, K.-F. Lee, and H.-W. Hon, "On semi-continuous hidden Markov modeling," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 689–692, 1990.

[47] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, pp. 164–171, 1970.

[48] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[49] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.

[50] C. Manfredi, M. D'Aniello, P. Bruscaglioni, and A. Ismaelli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices," *Medical Engineering & Physics*, vol. 22, no. 2, pp. 135–147, 2000.

[51] J. Harrington and S. Cassidy, *Techniques in speech acoustics*, vol. 8. Springer Science & Business Media, 1999.

[52] D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*. Regina: Department of Computer Science, University of Regina, 2003.

[53] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, no. 11, pp. 1477–1493, 1986.

[54] V. Parsa and D. G. Jamieson, "A comparison of high precision fo extraction algorithms for sustained vowels," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 1, pp. 112–126, 1999.

[55] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," in *Proceedings of the Symposium on Computer Processing Communications*, vol. 779, 1969.

[56] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *The Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 829–834, 1968.

[57] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *Signal Processing Magazine, IEEE*, vol. 21, no. 5, pp. 95–106, 2004.

[58] "Yin pitch estimator." `http://audition.ens.fr/adc/sw/yin.zip`. Online; accessed 27-02-2015.

[59] A. De Cheveigné and H. Kawahara, "Comparative evaluation of f0 estimation algorithms.," in *INTERSPEECH*, pp. 2451–2454, 2001.

[60] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7815–7819, IEEE, 2013.

[61] A. von dem Knesebeck and U. Zölzer, "Comparison of pitch trackers for real-time guitar effects," in *Proc. 13th Int. Conf. Digital Audio Effects*, 2010.

[62] G. Várallyay, "The melody of crying," *International Journal of Pediatric Otorhinolaryngology*, vol. 71, no. 11, pp. 1699–1708, 2007.