Automatic Baby Cry Diary

BY

BAHARE NAIMIPOUR
B.S., University of Illinois at Chicago, 2006

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Roland Priemer, Chair and Advisor
Rashid Ansari
Vladimir Goncharoff

ACKNOWLEDGEMENTS

I would like to thank Dr. Roland Priemer for his patience and guidance through my graduate study, as both my academic advisor and chair of my dissertation committee. I would also like to take the opportunity to thank the members of my dissertation committee Dr. Rashid Ansari and Dr. Vladimir Goncharoff.

I am very grateful to Dr. Tracy Magee of Riley Child Development Center for allowing me to use the recordings she obtained for this research.

I would like to acknowledge that this project was partly supported by a National Institute of Health grant to Dr. Roland Priemer and Dr. Tracy Magee.

I would also like to thank Dr. Gabriella Cerrato for her additional support and guidance, as well as my parents, grandparents mamany and daddy, and aunt Feloria.

Finally I would like to thank my husband, Mohsen, for his patience and advice.

BN

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LPC	Linear Predictive Coefficients
ST	Short-Term
AR	Autoregressive
STE	Short-Term Energy
ZCR	Zero Crossing Rate
ANN	Artificial Neural Network
FFBP	Feed Forward Back Propagation
MFCC	Mel-Frequency Cepstrum Coefficient
VAD	Voice Activity Detection
ACD	Automatic Cry Diary
MCD	Manual Cry Diary

SUMMARY

Crying is an infant's earliest and most effective mode of communication. This communication process can be disrupted if cry characteristics or acoustic attributes associated with infant crying are abnormal. The standard of measurement in the study of infant cry abnormalities has been a written cry diary. Cry diaries produced in the home by parents produce inconsistent and unreliable data. In this work, baby-cry was recorded over a 24-hour period in the natural home environment and digitized for computer-based analysis. The various sounds that have comparable energy or overlap baby-cry both in time and frequency were included in the recording. Our goal was to identify all baby cry time segments in order to automate the generation of a baby-cry diary.

The creation of our cry diary began with the study of the basic principles of speech processing. This led us to developing a Neural Network that would best classify baby cry and non-cry in a natural setting using a feature vector that would be extracted from each overlapping short-term frame of all recordings. In the end our Automatic Baby Cry Diary model consists of a 12-D Feature Vector and a Feed Forward Back Propagation Neural Network. Based on these results that are presented, we believe our Automatic Baby Cry Diary shows strong promise in developing practical baby cry analysis tools to aid physicians in diagnosing baby diseases such as Colic.

1. INTRODUCTION

1.1 Statement of Problem

Crying is an infant's earliest and most effective mode of communication. This communication process can be disrupted if cry characteristics or acoustic attributes associated with baby crying are abnormal. The standard of measurement in the study of baby-cry abnormalities has been a written cry diary. Cry diaries produced in the home by parents produce inconsistent and unreliable data. Research now suggests that many parents perceive baby crying five times greater than actual crying time [1]. In this work, baby-cry was recorded over a 24-hour period in the natural home environment and digitized for computer-based analysis. The various sounds that have comparable energy or overlap baby-cry both in time and frequency were included in the recording. Our goal is to identify all baby-cry time segments in order to automate the generation of a baby-cry diary.

1.2 Significance of the Problem

Crying is an infant's way of communicating with the outside world. The baby-cry indicates a need or want and motivates the listener to respond. It is a bidirectional communication process that consists of the cry itself (cry characteristics) and the salience of the cry to the care giving environment (parent perception) [2]. This communication process can be disrupted in two ways. First, the cry characteristics or acoustic attributes associated with infant crying may be abnormal. Distinct cry characteristics have been identified in infants with a variety of medical diagnoses. Cry characteristics are acoustic attributes associated with baby crying. The second disruption to the communication process is that the parental perception of

the cry may be atypical, either over or under responsive. Parent perception of baby-cry, temperament and behavior has been shown to be influenced by personal characteristics such as anxiety, depression and social support. It has been found that anxious parents are more likely to report their child is difficult to sooth and is more distractible, and depressed mothers are less sensitive to their infants' behavioral cues such as crying.

One of the most important tools for analyzing abnormal communication processes historically is a written cry diary [3]. A written cry diary is obtained using the receiver's (parent) written record of their perception of how much or how long their infant cries in 24-hours [3]. The cry diary, first published by Barr and colleagues (1988) is assumed to measure the amount of time the infant spends in alert, crying, fussing, and sleeping behavior during a 24-hour period. Prior research has reported a modest correlation ($r=0.60$) to parental reports of baby crying in a 24-hour period and an audio taped recording of that same 24-hour period [3]. Cry diaries produced in the home by parents do not provide accurate information on the quantity of the cry [4]. Since cry diaries must be maintained over a long period, they can create a hardship for the parent, which could alter the crying communication process.

Research indicates that unique acoustical characteristics of the cry (such as duration and pitch) may influence parental perception that the crying is excessive [4, 5, 6, 7]. Personal characteristics may also contribute to parental perceptions of the cry. These factors create an abnormal bidirectional communication process that may result in current and future problems for both the infant and the care-giver. Fortunately, technological advances in computerized sound

analysis now offer opportunities to assess crying more thoroughly and accurately by extracting the naturally occurring baby-cry from extraneous, overlapping environmental sounds.

1.3 **Background**

Diagnostic listening dates back to the time of the ancient Greeks and Hippocrates, but its value was overlooked until the mid 1800's [8]. It was at that time that Charles Darwin analyzed diagnostic listening in reference to baby crying and screaming in detail using photographs and drawings to demonstrate various expressions of emotion [9]. Nearly a century after Darwin's work, the most comprehensive treatment of baby-cry was established by a Scandinavian research group led by Olé Wasz-Höckert [10]. Wasz-Höckert *et al.* [11] published their first findings on the cry analysis of pain, birth, pleasure, and hunger cries using sound spectrographs in 1962. In 1968, they proceeded to publish a detailed statistical analysis of cries of both healthy and abnormal infants [10]. This publication was so important that it became a reference for baby-cry researchers for over two decades. More advanced spectrographic techniques began in the 70's with Michelsson *et al.* exploring baby-cries with different diseases and abnormalities [12, 13, 14, 15, 16, 17, 18, 19]. Their findings show that some abnormal baby-cry characteristics vary significantly from those in normal baby-cry.

The mid 70's was a time of significant advances in signal processing theories and computer technology. In 1974 Tenold examined the fundamental frequency and cry spectra of full-term and premature baby-cry using cepstral and stationary analysis [20]. In turn Corwin and Golub published their findings on computer aided analysis using computer signal processing techniques in the early 80's [8, 21, 22]. Their work shows promise for detecting a range of

abnormalities through baby-cry analysis. Finally in the late 80's and 90's more sophisticated signal processing techniques like the Fast Fourier Transform were applied to normal and abnormal baby-cry studies [23, 24, 25]. All these studies along with others have tried to identify unique features in the baby-cry signal that would eventually enable us to automatically classify them using a computer in a clinical setting. Although valuable information about the baby's physical, emotional, and psychological state could be gained through parameters like facial expressions, sleep, and suction abilities; acoustic analysis remains an important research area probably because of the inexpensive hardware requirements and the non-invasive nature of audio analysis itself [26].

For the past twenty years, work published in the area of baby-cry processing became more focused on a specific application. Based on the application, attempts to try various combinations of features and classification methods to obtain the optimal combination for use in a clinical setting were made. Within all of the recent research in the area of baby-cry, the work done by Petroni and Malowany [27], Xie and Ward [28], and Manfredi [29] are among the most comprehensive and unique. Petroni and Malowany classify anger, fear, and pain cries of normal infants using mel-frequency cepstrum coefficients (MFCC) and energy as their main cry characteristics. They tried three neural networks for classification, where they state the best overall results were obtained using the feed forward network. Xie defines a new parameter that describes the 'level of distress' of the baby-cry sound based on parental perception, and uses hidden markov models for their classification with decent results. Manfredi displays estimates of parameters such as fundamental frequency (pitch) and vocal tract resonant frequencies (formants) for newborn infants comparing modified variations based on autoregressive models

and the cepstrum approach. Manfredi shows how both methods have their advantages, but results were hard to judge because no classification of baby-cry was involved in their study and only a few examples were presented to explain their method.

Very recently, a few studies were done on classification of normal and pathological cry, specifically cries of deaf babies or with asphyxia. Orozco and Garcia [30] classify normal and deaf baby-cries by extracting 16 linear predictive coefficients, and using the feed forward neural network as their classifier that was trained with the scaled conjugate gradient algorithm. Their results showed 86% classification accuracy. Hariharan *el al.* [31] tried classifying normal baby-cries and cries of babies with asphyxia using statistical features extracted from the cry short-term Fourier transform (STFT), and three different neural network classifiers. The classification methods proved more than 95% accurate. In two other studies with the same classification aim, Hariharan uses different neural network classifiers and the same STFT statistical features which resulted in slightly lower than 95% classification accuracy in one study [32]. The other study uses only the neural network that gave the best results in the previous two studies, and linear prediction cepstrum coefficients (LPCC's) for the cry characteristic that also gave good results [33]. Other studies with similar aims, features extracted, and classifiers as the studies just mentioned have been done with slight variations, but again comparable results [34, 35, 36, 37].

1.4 **Significance of the Study**

In our study, not only did we need to consider various cry characteristics for cry classification, we also needed to consider the role parameters like voice activity detection (VAD) play in real world scenario's (i.e. the home environment, in the car, on the street, etc.). VAD

needs to be considered because our sound recordings were done in the baby's natural living environment. This makes for a more realistic application of any tool designed to analyze natural baby-cry . This being said, the short-term signal energy and zero-crossing rate have long been used as simple acoustic features for VAD [38, 39].

All of the previous work on baby-cry analysis has been done on baby-cry samples (as opposed to whole cry recordings), typically induced, and always in quiet environments. But this requires manual monitoring or extraction of baby-cry that has been recorded in a quiet environment. Time wise, this is not clinically practical. Our work automatically identifies natural baby-cry in the home environment which is different from all of the cry research done thus far. Our work will help create a more efficient baby-cry diary for more accurate diagnoses of babies with certain conditions such as Colic in the comfort of their home.

The following chapter will discuss the theoretical framework involved in this study. It begins with the basics of how human speech is produced, followed by a description of its mathematical model. The cry characteristics used in this study are then presented and explained. The remainder of the chapter introduces our chosen classification method. Chapter three presents the data that was acquired for this study and the steps that were involved in its processing. Subsequently the results of our processed data, or automatic cry diary, are presented and analyzed. The last chapter summarizes our work and talks about some aspects that can be addressed in the future.

2. THEORY

2.1 Speech Processing Basics

Baby-cry processing is a subcategory of voice and speech processing. Like voice and speech processing, the study of baby-cry processing relies upon some knowledge of acoustics, linguistics, physiology, psychology, computer science, and engineering. In this chapter, we will describe the speech processing theory we practiced to automate baby-cry recognition.

2.1.1 The Speech Production System

Before going into the details of speech processing, it is necessary to understand the basics of how speech is produced. The schematic diagram given in Fig. 1 shows the human vocal system.

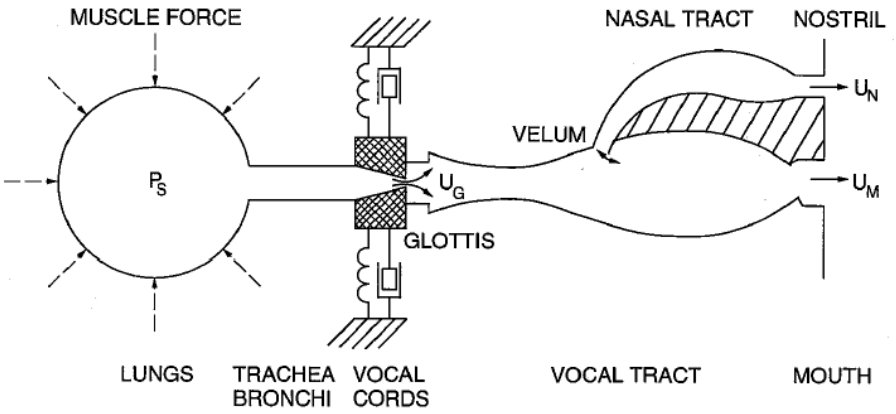

Figure 1: Schematic diagram of the human vocal system [40]

The driving force (excitation) for speech production comes from the lungs and trachea/bronchi. Air is expelled from the lungs and delivered to the vocal tract through the trachea/bronchi. From there the air is manipulated and enhanced (modulated) by the vocal tract or in some cases the nasal tract to produce speech sounds.

The vocal cords vibrate based on the air velocity, which depends on the sub glottal air pressure, in the lungs [41]. Vocal cord vibration disrupts airflow and as the cords open and close, the airflow breaks up into pulses as shown in Figure 2.

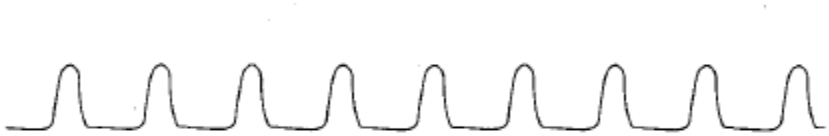

Figure 2: Typical glottal pulse train

The rate at which these pulses repeat is called pitch or Fundamental Frequency (F_0). These quasi-periodic pulses excite the vocal tract and are then modulated to produce voiced sounds. Unvoiced sounds on the other hand do not cause vocal cords to vibrate. Unvoiced sounds are produced by forcing air through a constriction in the vocal tract at a high enough velocity to produce turbulence. This action creates a broadband noise source that excites the vocal tract and produces unvoiced sounds.

The overall speech production system can be summarized with two functions, excitation and modulation. The excitation takes place mainly in the glottis, while modulation takes place in the various organs of the vocal tract [41]. From an acoustic standpoint, modulation takes place

primarily by means of filtering. For voiced sounds, glottal pulses (waveforms) can introduce a wide range of harmonics in a vocal tract. The vocal tract has its own natural frequencies, which are a function of its shape. These natural frequencies (formants) are the main modulating tool in the vocal tract that creates all the vowels and most of the consonants.

2.1.2 **Non-Stationary Nature of Speech Signals**

In the previous section we talked about how a speech signal is produced in the vocal tract system. In general, signals are characterized by many criteria such as periodic (as opposed to aperiodic), stationary (as opposed to non-stationary), harmonic (as opposed to non-harmonic), etc. One of the key distinctive characteristics of speech signals is non-stationarity. A stationary signal has frequency and spectral content that does not change with time. Since speech signals are non-stationary, their spectral content changes with time. Baby-cry, like speech in general, is non-stationary. Unlike conventional speech though, baby-cry is more harmonious. Therefore baby-cry is more short-term stationary than speech, a feature which has its advantages as will be discussed in the next section.

2.1.3 **Short-Term Processing**

In order to deal with the problem of baby-cry non-stationarity in computer analysis, we need to analyze baby-cry signals by dividing the signal into stationary or pseudo stationary time frames. We assume that cry features remain almost unchanged during each time frame. This way of processing signals is referred to as *short-term* processing. Baby-cry is processed over a sequence of short time intervals that may overlap. We assume that over each short-term interval the baby-cry signal is substantially stationary. In the figure below we have *N* samples of a signal

s(t) obtained at a sampling frequency rate f_s=1 / *T*, where *n* is the sample index number. Each short-term frame is $T_0 = MT$ seconds long. The number of time intervals (frames), N_0, depends on *M* and the amount of frame overlap.

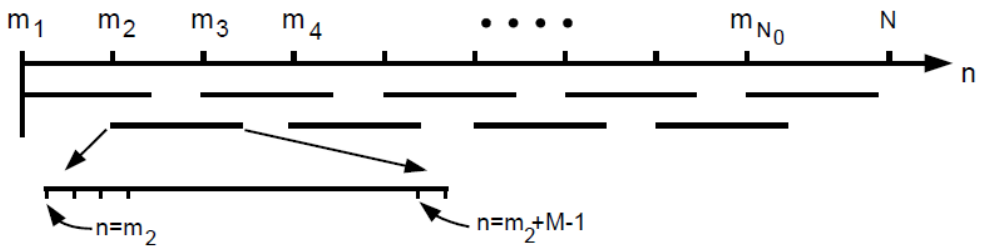

Figure 3: Segmentation of *N* samples into N_0 segments using a sliding window with overlaps

[42]

These time frames are processed using blocks referred to as 'windows'. The portion of the signal overlapped by each frame is windowed and sampled to obtain *M* samples of $s_p(t) = s(t)w(t)$, where $w(t)$ is a window function of duration T_0. The value of a finite signal *s(t)* during one short-term frame m_i, can be expressed mathematically as:

$$S'_n = S_n(m) = \sum_{n=0}^{M-1} w(n)s(m+n) \qquad (2.1)$$

where $m=m_i$, and $i=1,2,...,N_0$ is a time index that points to the beginning of a short-term frame as is illustrated in Figure 3. Equation (2.1) shows that only a part of *s(n)* that is overlapped by the window function is processed. The width of the short-term frame (*M*) depends on the degree of stationarity of our signal. Adult speech stationarity is 10-30 msec, therefore the value of *M*

should not exceed 30 msec for such applications. Baby-cry on the other hand is more stationary than adult speech, typically between 30-50 msec. The larger window size saves time which is important when processing cry diary's that are at least 24 hours long. There are various types of windows, but the ones most commonly used in the literature are the hamming, hanning, and rectangular windows.

2.1.4 **Correlation**

Correlation is a measure of similarity between two signals. Cross-correlation refers to the calculation of the similarity or correlation in signal characteristics between two different sequences. Auto-correlation refers to the cross-correlation of two different segments in the same sequence; but each segment occurring at a different time within the same overall sequence.

The autocorrelation $r(k)$ of an infinite stationary sequence $x(n)$ is given by:

$$r(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k) \tag{2.2}$$

The short-term autocorrelation, *R*, of a finite non-stationary sequence (e.g. baby-cry signal) *s(n)* is given by:

$$R_n(k) = \sum_{n=0}^{N-1-k} S'_n \, S'_{n+k} \tag{2.3}$$

where k is the autocorrelation lag and S'_n is defined in equation (2.1). Some of the important properties of the autocorrelation function include:

- It is an even function, meaning $r(k) = r(-k)$
- The value of $r(0)$ is equal to the energy if the signal is deterministic

2.1.5 **Linear Predictive Analysis**

Speech has been modeled as the result of a glottal pulse waveform convolved with the impulse response of the vocal tract as shown in Figure 4.

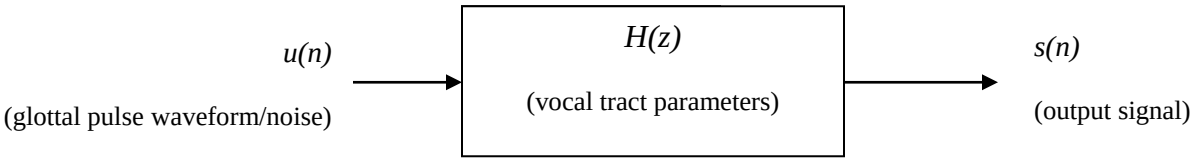

Figure 4: Speech production model

with unknown input *u(n)*. The output signal *s(n)* can be modeled such that:

$$s(n) = -\sum_{k=1}^{p} a_k\, s(n-k) + G \sum_{l=0}^{q} b_l\, u(n-l), \quad b_o = 1 \tag{2.4}$$

where a_k, $1 \le k \le p$, b_l, $1 \le l \le q$, and the gain G, are the parameters of the hypothesized system. This is a system where the output, *s(n)*, is a linear function of past outputs along with present and past inputs [43]. Therefore, this equation is a *Linear Prediction* of the speech signal *s(n)*. In speech processing, the most widely used vocal tract model is the all-pole or *autoregressive* (AR) model [41, 44, 28]. In the AR model we assume that *s(n)* is a linear combination of past outputs and the input *u(n)*:

$$s(n) = -\sum_{k=1}^{p} a_k\, s(n-k) + Gu(n) \tag{2.5}$$

This gives us the AR digital filter:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{2.6}$$

The value for G and the values for a_k can be estimated using *Linear Predictive Analysis*. A linear predictor with prediction coefficients α_k, order *p*, and at time *n* is defined as a system with the output:

$$\tilde{s}(n) = -\sum_{k=1}^{p} \alpha_k\, s(n-k) \tag{2.7}$$

The prediction error is defined as:

$$e(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^{p} \alpha_k\, s(n-k) \tag{2.8}$$

with the total squared error given by:

$$E = \sum_n e(n)^2 = \sum_n \left(s(n) + \sum_{k=1}^{p} \alpha_k\, s(n-k)\right)^2 \tag{2.9}$$

If we substitute equation (2.5) into (2.8), and our signal follows the model in Figure 4, then $e(n)=Gu(n)$ for $a_k=\alpha_k$.

We want to minimize the difference between the actual speech signal and its LP estimate assuming that $s(n)$ is a deterministic signal [43]. Given these assumptions, the total squared error E is given in equation (2.9). Since we would like to minimize the difference between the actual signal and its estimate, we minimize the total error by setting:

$$\frac{\partial E}{\partial \alpha_i} = 0, \qquad 1 \le i \le p \tag{2.10}$$

Using equation (2.10) and the equation for the total squared error, we obtain:

$$\sum_{k=1}^{p} \alpha_k \sum_n s(n-k)s(n-i) = -\sum_n s(n)s(n-i)\,, \qquad 1 \le i \le p \tag{2.11}$$

In order to estimate the values with equation (2.11), we use the *Autocorrelation Method*. The Autocorrelation method has been shown to be effective in speech processing [45, 46]. Going back to the discussion of autocorrelation in the previous section, equation (2.11) can be simplified to:

$$\sum_{k=1}^{p} \alpha_k R(i-k) = -r(i), \qquad 1 \le i \le p \tag{2.12}$$

where $r(i) = [r(0), r(1), \dots, r(p)]^{\mathrm{T}}$ is the *autocorrelation vector*, $\alpha_k = [\alpha_1\,, \alpha_2, \dots, \alpha_p]^{\mathrm{T}}$ is the *filter coefficient vector*, and $R(i-k) = \begin{bmatrix} r(0) & \dots & r(p-1) \\ \vdots & r(0) & \vdots \\ r(p-1) & \dots & r(0) \end{bmatrix}$ is the *symmetric toeplitz autocorrelation matrix*. We can now solve for the LP filter coefficients.

In many areas of speech processing, LPC's have shown excellent results with the development of efficient LPC calculating algorithms [41, 44]. One of the many important speech parameters that an LPC algorithm helps us to obtain are cepstrum coefficients, which will be further discussed in Section 2.2.

2.2 Cry Pattern Recognition

Short-term processing was discussed in the previous section as a means of being able to process a non-stationary speech signal. In our research, we want to process signals in order to automatically classify every second of a recording as either baby-cry or non-cry. In order to do so, we need to find a set of features that have proven to best describe baby-cry. Speech recognition has been a subject of study for quite some time, as opposed to baby-cry recognition, which has a more recent history. Since both adult speech and baby-cry recognition are similar in origin, the features studied and tested in baby-cry pattern recognition stem from those used in adult speech recognition. From here on we will replace the use of the term 'speech' with 'baby cry' for the sake of consistency and relevance to our specific work. The next subsections will discuss feature selection, and then later a more detailed discussion of the features used in our study is given.

2.2.1 Feature Selection

The data inputted to a baby-cry recognizer consists of both relevant and irrelevant information. Features play the important role of minimizing or ignoring the irrelevant information while maximizing or highlighting the relevant information that is presented to the

classifier. The classifier then sorts the information described by the features into designated classes. Ideally, although generally impossible, we would like our features to [41]:

1. Vary widely from class to class
2. Be insensitive to extraneous variables (environment, recorder noise, etc.)
3. Be stable over long periods of time
4. Be frequently occurring
5. Be easy to measure
6. Not be correlated with other features

In a more quantitative description, we should aim to select features leading to a large between class distance and small within class variance in the *feature vector* space [47]. This means that if we have N extracted features for each short-term window, these features are all set alongside each other creating a feature vector x for that specific short term window:

x = [feature 1 value, feature 2 value, feature 3 value,..., feature N value]

This vector is a point in an N-dimension *feature space*. Now, assume that all of the short-term feature vector points are plotted in this N-dimension feature space. In order to classify these vector points into our desired classes, a pattern needs to be found. This can be done if we put up dividing walls in the feature space for each desired class. Thus, each class is defined by a set of dividing walls or *decision boundaries*. Once these decision boundaries (hyperplanes) are defined for each class, our classification problem is solved.

One big problem in studies that involve feature extraction, is the huge number of features to choose from. Trying to find the optimal features for a specific application is a difficult task. Therefore, some features are often chosen intuitively. Looking at the results of related research

and having some knowledge of anatomy, phonetics, etc. helps build a better intuition of which features might work well for a specific application.

2.2.2 **Features used in this study**

The features used in our study were chosen in the following manner. Initially we looked at frequently used features in speech processing. Then, from those features, we chose speech processing features that were also applied in various baby-cry research that has shown promising results [28, 34, 35, 30, 36, 33].

Short-Term Energy (STE):

The energy associated with baby-cry is time varying in nature. The interest of any automatic processing of baby-cry, is to know how the energy varies within a short term region of the signal. As mentioned before, acoustic signals generally consist of voiced, unvoiced and silence regions. The energy associated with the voiced region is large compared to the unvoiced region. The silence region will have negligible energy. Short-Term Energy is defined as:

$$e(m) = \sum_{n=0}^{M-1} (s(n)w(m+n))^2 \tag{2.13}$$

where $s(n)$ is the recorded signal, $w(n)$ is the window function, and the remaining parameters are defined following equation (2.1).

Short-Term Zero Crossing Rate (ST ZCR):

The Zero Crossing gives information about the number of zero-crossings present in a given signal or in other words, the number of times at which the sign of the signal changes from positive to negative and vice versa. Intuitively, if the number of zero crossings is large in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information. Similarly, if the number of zero crossings is small, the signal is thought to be changing slowly and may contain low frequency information. The ZCR gives us indirect information about the frequency content of the signal. For a stationary signal $x(n)$, the total number of zero crossings is defined as:

$$z = (\frac{1}{2}) \sum_{n=0}^{\infty} |sgn(x(n)) - sgn(x(n-1))| \tag{2.14}$$

where

$$sgn(x(n)) = \begin{cases} 1, & if\ x(n) \geq 0 \\ -1, & if\ x(n) < 0 \end{cases}$$

With the stationary case in mind, the ST ZCR of *s(n)* can be expressed as:

$$zcr(m) = \frac{1}{2M} \sum_{n=0}^{M-1} |sgn(s(n)) - sgn(s(n-1))| w(m+n) \tag{2.15}$$

Again all parameters are defined following equation (2.1).

Cepstral Coefficients:

LPC derived cepstral coefficients have shown best recognition accuracy in automatic speaker recognition [48], and they are commonly used in recent baby-cry recognition research [28, 35, 33]. Cepstral coefficients give high speaker recognition accuracy, and are invariant to fixed linear spectral distortions from recording and transmission environments [48]. Results have been shown to be similar whether the Cepstral Coefficients were computed with Fourier Tranforms or with LPC's, but the LPC method is twice as fast [49]. Once the filter coefficients for the Linear Predictive All-Pole model for a frame of the baby-cry signal using the Autocorrelation Method are calculated, the Cepstrum Coefficients are computed as follows:

$$\begin{cases} c_1 = \alpha_1 \\ c_m = \alpha_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k \alpha_{m-k} & 1 < m < p \end{cases} \tag{2.16}$$

where p is the order of the predictor, α_i is the i^{th} linear predictor coefficient, and c_i is the i^{th} Cepstrum Coefficient.

An LPC in and of itself is not very useful because it is influenced by the frequency response of the recording device and the transmission system. Cepstral Coefficients on the other hand, have the advantage of presenting a set of parameters that are invariant to any fixed frequency response distortions. The mean value for each coefficient over time is subtracted from each coefficient function, which yields a signal that minimizes environmental and intra speaker effects [48].

In this section features and the reason for their extraction was discussed. In the next section we will discuss how to use these feature vectors to discover patterns that will allow us to distinguish baby-cry from other sounds.

2.3 Cry Pattern Classification

In previous cry research, various computer classification methods were used to help distinguish cry from non cry, normal from various pathological cries, or different baby sounds. The more commonly used methods are the Hidden Markov Model [28, 50, 51] and Neural Networks [35, 30, 36, 37, 31, 33, 32, 27] . In our research, Neural Networks were used for 'cry' and 'non-cry' classification of the recordings on a second by second basis.

Neural Networks came from the idea to systematically model the workings of the human nervous system. In practice, Neural Networks consist of a set of nodes and a set of links. The nodes correspond to neurons and the links correspond to the data flow between the neurons. Through these links, a series of actions are applied to the inputs of nodes based on our desired output.

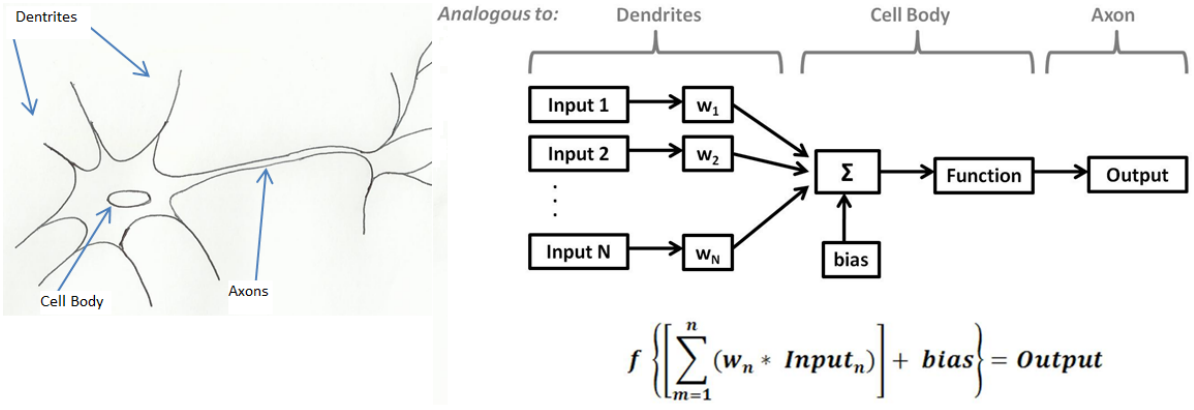

Figure 5: An artificial neural network analogous to a biological neuron

An analogy of the mathematical model and the biological system is shown in Figure 5. As can be seen, these actions consist of weights, biases, summations, and a function applied to the output of a node, based on our desired output. Weights and biases are generally scalars, and the function applied generally depends on the system being modeled and the chosen training method to be implemented. The neurons connected in a layered network are configured so that the output of every neuron in a layer is inputted to every neuron in the consequent layer.

Before choosing the training method by which we want to train our Neural Network, we need to determine its learning path. A Neural Network can either learn with supervision, known as *Supervised Learning*, or without supervision, known as *Unsupervised Learning*. Supervised learning occurs when the Neural Network is trained with an input training data set and a desired output target vector. The Neural Network adjusts its weights and biases until the error is minimized between target output and the network output, given the input training data. Unsupervised learning refers to allowing the network to adjust itself given only the input training data, but without a target output vector. Unsupervised learning is a type of machine learning algorithm used to draw connections between input data without labeled responses. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. Machine learning algorithms can adaptively improve their performance as you increase the number of samples available for learning. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. These clusters are formed based on a measure of similarity defined by metrics such as Euclidean or probabilistic distance.

In this study, we want to automatically classify every second of recorded data as cry or non-cry sound. All recordings were individually listened to by a person that was not the baby's parent in order to avoid bias; and each second was classified as cry or non-cry sound. This created a Manual Cry Diary which is the target reference for our work. With the existence of the Manual Cry Diary, Supervised learning was chosen as our learning path.

There are a variety of methods used to train Neural Networks using Supervised learning. The difference between these methods is in the technique they use to update their weights and biases to minimize the error between network output and desired output. These methods have been studied in various references [52], along with a summary of many of these methods applied in cry research [35, 30, 36, 37, 31, 33, 32, 27]. The most promising and frequently used Supervised learning method used in cry research using Neural Networks is Feed Forward Back Propagation. The next section will discuss the mathematical details of the Feed Forward Back Propagation method.

2.3.1 **Feed Forward Back Propagation**

In the Feed Forward Back Propagation (FFBP) method, initially small random weight and bias input values are set and an output vector is calculated based on the input vector. In this step since we start from an input vector and end with an output vector, the system is feeding forward. In the next step the error between the target output and system output is calculated. This error is then back propagated from the output node to the input node. The weights and biases are updated based on the error at each node. The error is recalculated and this process is repeated until the error is less than a given threshold or the maximum number of epochs (the number of

times the system adjusts its weights and biases from input to output after error calculation) is reached.

The weights and biases are generally updated based on the gradient decent optimization approach. In order to find the optimal weight and bias values using gradient descent, back propagation takes steps proportional to the negative of the gradient of the error surface at the current point. This method leads towards a global error minimum along the steepest vector of the error surface. Since the nature of this error space cannot be known beforehand, neural network analysis often requires a large number of training runs to determine the optimal solution.

The FFBP algorithm can be mathematically summarized as follows:

$$a^0 = input \tag{2.17}$$

$$a^m = f^m(W^m . a^{m-1} + b^m), \quad m = 1,2, \dots, M \tag{2.18}$$

Equations (2.17-2.18) summarize the Feed Forward step of the FFBP algorithm. The index m indicates the layer number meaning that $m = M$ is the output layer number. The function f is the network activation/transfer function. Neurons can use any differentiable activation function to generate their output. The vector a^m is the output vector of the mth layer, while W and b are the weight and bias vectors respectively.

In the error Back Propagation step, the error in the output vector is calculated as follows:

$$\varepsilon^M = (t^M - a^M) . f'(a^{M-1}) \tag{2.19}$$

where t is our target vector, and f' is the derivative of the network activation function. The network activation function for FFBP is usually the sigmoid or hyperbolic tangent function by default because these functions have derivatives that can easily be calculated.

The error vector for all other layers except the output layer is then calculated with the following relation:

$$\varepsilon^m = (\sum \varepsilon^{m+1}.W^{m+1}) f'(a^{m-1}), \quad m = M-1, M-2, \dots, 1 \tag{2.20}$$

Using the calculated values obtained in equations (2.17 - 2.20), the updated weights and biases are calculated as follows:

$$W_{new}{}^m = W^m + \gamma.\varepsilon^m.a^{m-1}, \quad m = 1,2, \dots, M \tag{2.21}$$

$$b_{new}{}^m = b^m + \gamma.\varepsilon^m, \quad m = 1,2, \dots, M \tag{2.22}$$

where γ is the learning rate. The learning rate applies a greater or lesser fraction of the respective adjustment to the old weight. If this factor is set to a large value, then the neural network may learn more quickly requiring less training runs, but if there is a large variability in the input set then the algorithm may oscillate and become unstable. On the other hand, if the learning rate is too small the algorithm will take too long to converge. Therefore it is not realistic to expect to find the optimal setting for the learning rate before training. Often the better solution is to have an adaptive learning rate that attempts to keep the learning rate as large as possible while keeping learning stable by changing the learning rate based on the complexity of the local error surface.

An adaptive learning rate attempts to keep the learning step size as large as possible while keeping learning stable. It does so by calculating new weights and biases at each epoch

based on the current learning rate. New outputs and errors are then calculated and if the new error is greater than the old error by more than a predefined ratio, the new weights and biases are discarded and the learning rate is decreased. If the new error is less than the old error, the learning rate is increased by multiplying the learning rate by a predefined ratio. This procedure increases the learning rate, but only to the extent that the network can learn without large error increases resulting in a near-optimal learning rate for the local terrain.

Once a Neural Network is 'trained' to a desirable level it may be used as a classification tool on other data. Hence the user no longer specifies any training runs and instead the newly trained network only needs to forward propagate. New inputs are presented to our customized network where they are filtered and processed by the middle layers as if training were taking place, however at this stage the output is retained and no back propagation occurs [53].

MATLAB version R2009 was used throughout this project. The Neural Network Toolbox in MATLAB was used to implement the Neural Network described above. The *newff* function creates the FFBP network, based on the parameters we choose. Our FFBP NN has 12 neurons, and one hidden layer, making a total of 2 layers including the output layer. The transfer function of each layer can also be individually chosen. In our NN we chose the 'tansig' function for the hidden layer and 'purelin' for the output layer. Our NN training function is set to use the Gradient Decent with Adaptive Learning Rate ('traingda'). Another useful parameter is the number of epochs. An *epoch* corresponds to the entire training set going through the entire network once. We set the number of epochs to 200, therefore the weights and biases of our NN can be updated up to 200 times or until our set error goal of 5e-2 is achieved when tested on the training set.

3. RESULTS

3.1 Data Acquisition

The recordings used in this project were obtained by Dr. Tracy Magee of the College of Nursing. The procedure involved in obtaining these recordings included recruiting participants based on a series of requirements, teaching each participants' guardian the proper use of the recording system, and obtaining manual cry diaries after the 24-hour recordings were completed. The following sub sections offer greater details regarding the aforementioned.

3.1.1 Research Participant Requirements and Recruitment

Before any data can be recorded, participants had to be recruited and evaluated to fit the study needs and requirements. The inclusion criteria to participate in this research study included: 1) The infant needs to be born full term, and at the time of recording be between the ages of 4 weeks and 16 weeks 2) mother must be at least 18 years of age; and 3) mother must be able to read and write English. Exclusion criteria included: 1) the parent is not the legal guardian, 2) active family involvement of the Illinois Department of Children and Family Services (DCFS), and 3) known drug use in the home. Infants were not excluded for low birth weight. Parents were not excluded for known medical conditions. Only English speaking mothers were recruited for this feasibility study for two reasons: 1) the principal investigator (PI) is English speaking and to understand the challenges and barriers to this research, accurate and timely communication is critical; and 2) the tools being used are not all readily available in other languages and translation of the tools is timely and costly and would not advance the aims of this methodology study.

Research participant recruitment was conducted in the outpatient pediatric primary care center and an outpatient women's health center of a large urban, Midwestern medical center, and the Fussy Baby Network. Approximately 60 newborns per month are seen in the primary care pediatric clinic. All of the recruitment sites serve large, urban areas that reflect the racial/ethnic composition of Chicago as reported in the 2010 census: Hispanic (29%), Caucasian (31%), African American (33%), and Other (7%). Infants referred to the Fussy Baby Network routinely have a complete medical work up in an attempt to find a medical reason for the crying. This workup most often includes: complete physical examination, dietary alterations, and often a trial of medication for gastroesophogeal reflux (GER). Approximately 30 infants and their families are enrolled monthly at the Fussy Baby Network. Once the research participants were recruited and passed the inclusion criteria, they were individually trained in their home setting on the use of the recording system.

3.1.2 **The Recording System**

A small commercially available recorder was used to acquire infant vocalizations for 24 hours in the home and was able to provide recordings of sufficient quality for developing an automated analysis of crying. The recording system (Advanced Security Products model number MBR 565) is a small, battery (two AAA cells) operated, digital voice recorder that is completely self-contained. In voice mode it has a bandwidth of 4 KHz. The recorder has a 2 GByte memory, giving it a capacity to hold 116 hours of super high quality (SHQ) voice recordings. The recordings are uploaded to a PC via the recorder's USB port, and then they are converted to MP3 format. The size of the recorder is 3-7/8" tall x 1-1/4" wide x 5/8" thick. Each family received

two recorders. Even though the recording system can record for more than 24-hours, the battery life is 12 hours, requiring two recording systems per family.

These tiny portable recording systems were housed in a vocal recording vest system (VRV), to record naturally occurring infant vocalization including in-the-home sounds. The VRV was found acceptable to parents, did not interfere with the daily activities of the mother, the infant, or with mother-infant interactions.

3.1.3 **The Manual Cry Diary**

Research suggests that normal infants between the ages of 4 weeks and 16 weeks cry on average 1-2 hours per day. Given this, we anticipated 24 hours of recording will result in approximately 60-120 minutes of crying per infant. Using 24-hour cry data collected in the homes with the VRV from 7 normal infants, manual cry diaries (reliable master) were obtained that reflect the number of cry events and number and duration of cry episodes. Two trained and blinded research assistants (RA) listened to and transcribed sound data from the 24-hour recordings collected in the homes and created a manual cry diary that served as the master from which to establish the reliability of the computer generated cry diary. Only crying sound was transcribed. In the manual cry diary, the RA recorded the number of cry events (a cry lasting at least one second). By comparing the number of occurrences of cry events generated by the two methods (automated and manual cry diary) we will validate the accuracy of the baby-cry data extraction algorithm.

3.2 **Data Processing**

The 24-hour recordings were digitized and converted into wav format using Switch Sound File Converter, with a sampling rate of 8000 samples per second and 16 bit resolution. Wav file recordings were then time gated and separated into individual wav files, using Direct MP3-WAV Splitter software for two purposes. Direct MP3-WAV Splitter was used to separate various 1 minute cry and non-cry samples to train the Neural Network in recognizing cry from non-cry sound. Cry recordings also needed to be divided into 5 minute wav files so they could be processed in MATLAB. MATLAB was found to run out of memory when wav files were longer than 5 minutes long. Therefore the recordings were split into 5 minute segments and processed one by one in MATLAB for each 24-hour recording.

In our algorithm, MATLAB processes each segment and then categorizes each second as ‘cry’ or ‘non-cry’ based on a list of criteria. The output of the program is a series of 1’s (cry) and 0’s (no cry) known as the ‘cry vector’ for every second of the 24-hour recordings. The duration of a cry episode and/or all the cry’s in the recording can then be easily determined from this cry vector. A graphical explanation is shown below:

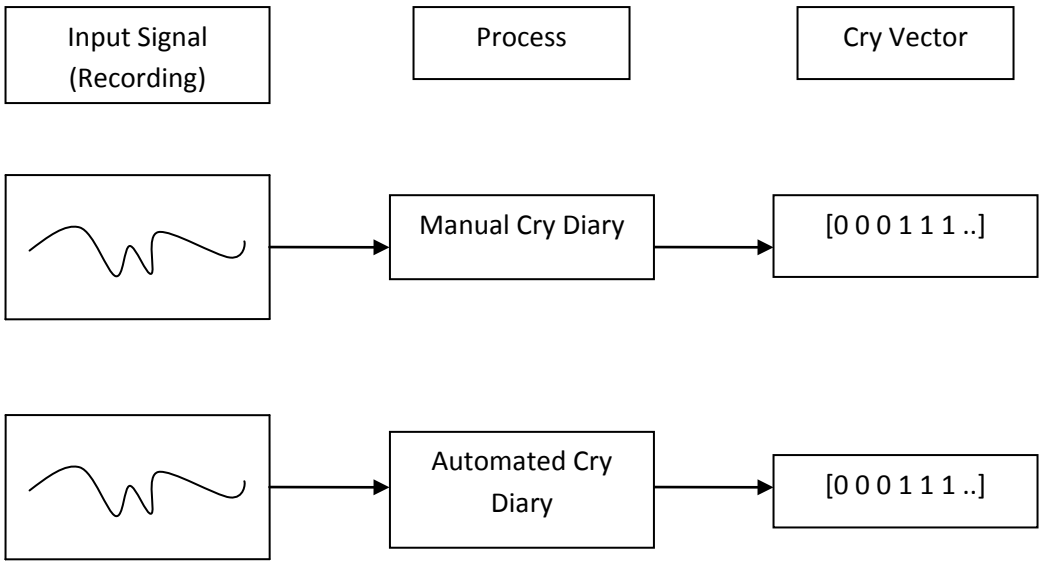

Figure 6: Comparison of Manual vs. Automatic Cry Diary using cry vectors

Once the Manual and Automatic Cry Diary Vectors are obtained, they can be compared in order to determine the efficiency of the algorithm used to obtain the Automatic Cry Diary.

Most of our algorithm has been explained in one form or another in Sections 2.2.2 and 2.3.1. The overall process of Automatic Baby Cry Diary is summarized in Figure 7.

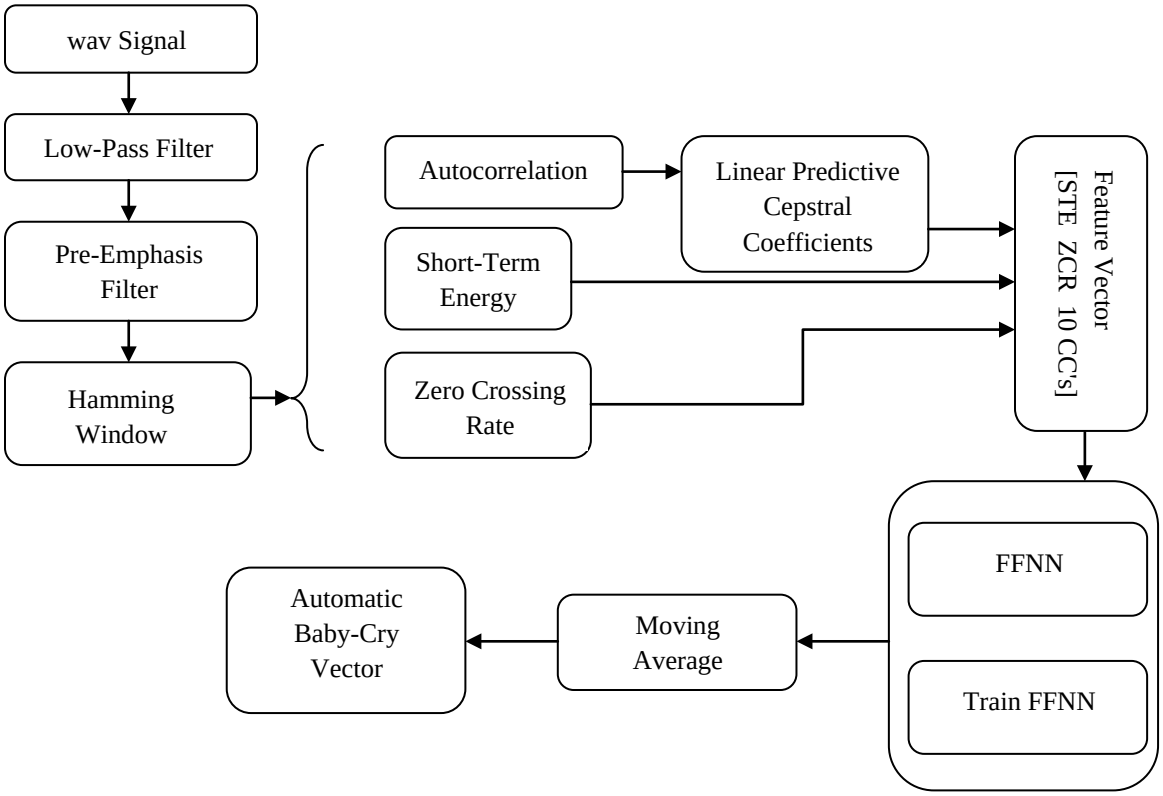

Figure 7: Graphical Description of Automatic Baby Cry Diary Program

After the raw recordings are formatted and digitized, the signal is low-pass filtered at 3 kHz in order to smooth the signal. A pre-emphasis filter with a transfer function of $1\text{-}0.95z^{-1}$ was also used prior to analyzing the signal in order to remove dc components and relatively flatten the spectrum so as to reduce the effect of the glottal waveform and lip radiation characteristics [28]. After low-pass filtering and applying the pre-emphasis filter, the signal needs to be windowed for analysis. The hamming and rectangular windows are mentioned in many speech processing texts. The hamming window is probably the most commonly used in speech applications because the rectangular window gives maximum sharpness but has large side-lobes

(ripples) in the frequency domain, while the hamming window blurs in frequency but produces much less side-lobe leakage. The large side-lobes in the rectangular window have been shown to mask higher formants when used to calculate common speech features like autocorrelation [41]. In our work, a sliding hamming window is used with a width of 32 ms and an overlap (sliding step) of 10 ms.

After filtering and windowing the signal, we extracted the 12 features used in this study that were described in Section 2.2.2. So for every window (time frame), we have a 12-dimension feature vector describing that window based on its energy, zero-crossing rate, and its first 10 cepstrum coefficients. The neural network is trained using various 1 minute samples of cry and non-cry from our recordings. Each 1 minute sample is set as 1 if it is 1 minute of cry and 0 if it is a non-cry sample in the Target vector. The Neural Network is then trained based on the feature vector associated with each sample and its respective Target value. Once this step is completed, the entire cry recording is ready to be classified as cry or non-cry for every window using the newly created Neural Network. The NN classifies every sliding window as cry or non-cry, while the Manual Cry Diary is classified as cry or non-cry every second. The NN also classifies each frame independently of the frames directly before or after it. The Manual Cry Diary (MCD) is a good tool in detecting the cry episodes, because a trained listener is aware of when a cry episode begins and ends regardless of minor breaks in the cry continuity. On the other hand the NN does a much better job of classifying individual cry events. So we needed to post-process the FFNN results in order to solve these issues in order to compare the Manual and Automatic Cry Diary (ACD) vectors for all the recordings. A Moving Average Filter is a surprisingly simple technique that solved a major portion of these issues, allowing us to use the FFNN results as a method that

identifies cry episodes much better than just individual cry events. For every frame, the Moving Average Filter averages the classification results of a 2 second span (1 second prior and one second after). Then depending on whether or not the average of a frame is above or below our tested threshold of 0.45, the new classification of that frame is set as 1 or 0 respectively and we can now move from the frame realm to the second realm . We can now move from the frame realm to the second based realm and are ready to compare the Manual and Automatic Cry Diary.

3.3 **Tabulated Results**

Results were initially evaluated based on individual recordings. So each recording had a custom FFNN that was trained using samples from that individual recording. The results summarized in the Table 1 below show that our approach is promising.

Recording Name	Total Recording Time (seconds)	Error (seconds)	Correlation between ACD and MCD
MNRS 2	64800	702	98.9%
MNRS 3	52200	2371	95.5%
MNRS 4	53700	755	98.6%
MNRS 5	86400	1181	98.6%
MNRS 7	86400	2533	97.1%
MNRS 9	86400	5895	93.2%

MNRS 10	86400	7394	91.4%

Table 1: Individual Recording Results

Our next step in training and testing a Neural Network that could be used in our Automatic Cry Diary involved training a new NN based on training sets from multiple recordings. Then we looked at the classification results when this new NN was used in generating an Automatic Cry Diary for recordings that had no samples included in training set of the NN. We stopped adding training samples from new recordings once our results changed drastically as a result of over fitting. Ultimately recordings MNRS 7, 4, and 3, were involved in training the Automatic Cry Diary's FFNN. The classification results are shown in Table 2.

Recording Name	Total Recording Time (seconds)	Error (seconds)	Correlation between ACD and MCD
MNRS 2	64800	1215	98.1%
MNRS 5	86400	1178	98.6%
MNRS 9	86400	2072	97.6%
MNRS 10	86400	2692	96.9%

Table 2: Automatic Baby Cry Diary Results

3.4 **Analysis**

The results observed in Table 2 give us assurance that the Automatic Baby Cry Diary can successfully estimate the length of time a baby cry's in a natural setting. The Correlation is simply the Error divided by the Total Recording Time. The Error is calculated by summing the difference, in seconds, between the Automatic Cry Diary and Manual Cry Diary for every five minutes of recorded time. This made it much easier to determine not only cry events, but correctly identified cry episodes. It enabled us to troubleshoot problem time ranges rather than individual second by second errors. Many of the error seconds were correctly identified as non cry or cry, because although the Manual Cry Diary is supposed to identify every cry event, many of the identifications are based on a continuous range of seconds being identified as cry purely because they are part of a cry episode. Baby fussing was also identified as cry in many instances of the Manual Cry Diary, which was not identified as such in the Automatic Cry Diary. What could not be changed in this project was the way the Manual Cry Diary was written, but what could be changed was the way the Automatic Cry Diary should identify cry. The Automatic Cry Diary needed to work more like a human, and it was from this analysis that we reached several post processing approaches before succeeding with the Moving Average Filter approach in the post processing phase. For four completely different babies in completely different settings, our Automatic Baby Cry Diary has matched the Manual Cry Diary by up to 98.6%, with the worst correlation being 96.9%. Therefore the worst correlation is still almost 2% better than our target of 95% correlation.

4. CONCLUSION

4.1 Summary

Until now, the standard of measurement in the study of baby cry abnormalities has been a written cry diary. The problem is that cry diaries produced in the home by parents produce inconsistent and unreliable data. Research has shown that many parents perceive baby crying five times greater than actual crying time. Parent perception of baby-cry, temperament and behavior has been shown to be influenced by personal characteristics such as anxiety, depression and social support. In this study our goal was to propose a possible solution for this problem by automating the generation of a baby-cry diary.

For the past twenty years, work published in the area of baby-cry processing became more focused on a specific application. Based on the application, attempts to try various combinations of features and classification methods to obtain the optimal combination for use in a clinical setting were made. This being said, all of the previous work on baby-cry analysis has been done on baby-cry samples (as opposed to whole cry recordings), typically induced, and always in quiet environments. But this requires manual monitoring or extraction of baby-cry that has been recorded in a quiet environment. Time wise, this is not clinically practical.

Our recordings were done over a 24-hour period in the natural home environment. The various sounds that have comparable energy or overlap baby-cry both in time and frequency were included in the recordings. Our work automatically identifies natural baby-cry in the home environment which is different from all of the cry research done thus far. Our work will help

create a more efficient baby-cry diary for more accurate diagnoses of babies with certain conditions such as Colic in the comfort of their home.

The creation of our cry diary began with the study of the basic principles of speech processing. This led us to developing a Neural Network that would best classify baby-cry and non-cry in a natural setting using features that would be extracted from each ST frame. In the end our Automatic Baby Cry Diary model consists of a 12-D Feature Vector and a FFBP NN. The 12 features include ST Energy, ST Zero Crossing Rate, and the first 10 Cepstrum Coefficients. The NN weights and bias values are updated using GDA optimization. Our NN was trained with cry and non cry samples of three of our 24 hour recordings. Our ACD has been tested on four of the 24-hour baby recordings with ACD and MCD correlating up to 98.6%, with the lowest being correlation being 96.9%.

Based on these results, we believe our Automatic Baby Cry Diary shows strong promise in developing practical baby-cry analysis tools to aid physicians in diagnosing baby diseases such as Colic. It is important to note however, that different baby diseases may have different cry patterns that may call for slightly modified parameters to better identify each individual disease. Therefore a universally valid ACD may be difficult to define, but the methodology and concepts established in this research, are universally applicable.

4.2 **Recommendations for Future Work**

Although this project has ended with promise, there were a few problems along the way that can be avoided in future studies. The most prominent of these issues was having to work

with recordings and Manual Cry Diaries taken and written by someone unfamiliar to the engineering field. In the future, I would recommend the engineer working with the data or at least familiar with basic signal processing research take the recordings. Also, the engineer and person involved in writing the Manual Cry Diary should clearly define what is cry and what isn't before writing the Manual Cry Diaries. Many of the seconds were incorrectly recorded in the Manual Cry Diary because fuss and any type of noise the baby made that possibly signaled discomfort was recorded as cry. Other times many seconds were classified as cry in the Manual Cry Diary maybe because they seemed to be part of the same cry episode, while they were physically not cry. This made it very difficult to define what is cry and non cry from a computer standpoint, and the results were not good during such time periods.

Another problem, which caused bad results, occurred when caregivers would sing to the baby. It might be hard to believe, but baby-cry is harmonic, just like singing. This can be clearly seen from looking at a simple spectrogram. So when we are looking at the frequency properties of a normal cry, we see that over time the cry smoothly moves from one frequency to the next without discontinuities. For this reason, we believe that the computer would mistake singing with crying in the Automatic Cry Diary. Therefore for future recordings, it would be a good idea to ask the guardian to avoid singing during the 24-hour recording period.

One other interesting problem we noticed, was that the NN is trained to classify cry in general, regardless of whether this is the cry of one baby, or many babies in the same recording. A person can tell the difference between one baby's cry from another baby's cry, but the Automatic Cry Diary is trained to identify any baby-cry. This created some bad correlation

results in one of the recordings, where there was another child in the recording whose cry was different from the baby of interest, but was nonetheless still cry. Therefore for future recordings in order to avoid this problem, we recommend only the baby of interest be heard in the recording.

In the future we believe this work can help the diagnosis of baby diseases like Colic. Colic is currently mainly identified using a parent's written diary and based on the number of hours a baby cries in a day, which is more than three hours if the baby has Colic. The Automatic Cry Diary will help not only identify Colic based on the amount of cry, but also Colic cry characteristics, which many caregivers are unfamiliar with and not considered in the written cry diary. Another direction this research might be helpful in, is in identifying a baby based on his/her cry when there is more than one baby in the same environment.

We have presented a good foundation for this work that has presented good results for normal cries. We would like to optimize this algorithm using other NN methods, varying the NN parameters, and adding or evolving features with the hope of even better, more robust results for varying applications. One of the other features we would like to use is the Mel Frequency Cepstrum Coefficient instead of Cepstrum Coefficients. Other good features worth testing are sound quality metrics that are very commonly used to describe mainly industrial machinery sounds.

CITED LITERATURE

[1] I. St. James-Roberts, "Infant crying and its impact on parents. In New Evidence on Unexplained Early Infant Crying: Its Origins, and Nature and Management," Johnson & Johnson Pediatric Institutes, Skillman, New Jersey, 2001.

[2] P. S. Zeskind, "Impact of the Cry of the Infant at Risk on Psychosocial Development," Encyclopedia on Early Childhood Development: Centre of Excellence for Early Childhood Development, Montreal, Quebec, Canada, 2005.

[3] R. Barr, M. Kramer, C. Boisjoly, L. McVey-White and I. Pless, "Parental diary of infant cry and fuss behaviour," *Archives of Disease in Childhood,* vol. 63, pp. 380-387, 1988.

[4] L. LaGasse, A. Neal and B. M. Lester, "Assessment of infant cry: acoustic cry analysis and parental perception," *Mental Retardation and Developmental Disability Research Review,* vol. 11, pp. 83-93, 2005.

[5] P. Zeskind and R. Barr, "Acoustic characteristics of naturally occurring cries of infants with "colic".," *Child Development,* vol. 68, pp. 394-403, 1997.

[6] B. Lester, C. Boukydis, C. Garcia-Coll and M. Peucker, "Developmental outcome as a function of the goodness of fit between the infant's cry characteristics and the mother's perception of her infant's cry," *Pediatrics,* vol. 95, no. 4, pp. 516-521, 1995.

[7] B. Fuller, M. Keefe and M. Curtin, "Acoustic analysis of cries from "normal" and "irritable" infants... including commentary by Garvin BJ with author response.," *Western Journal of Nursing Research,* vol. 16, pp. 243-253, 1994.

[8] H. L. Golub and M. J. Corwin, "A Physioacoustic Model of the Infant Cry," in *Infant Crying*, New York, NewYork, Plenum Press, 1985.

[9] C. Darwin, The Expression of the Emotions in Man and Animals, London: J. Murray, 1872.

[10] O. Wasz-Höckert, J. Lind, V. Vuorenkoski, T. Partanen and E. Valanne, "The Infant Cry: A Spectrographic and Auditory Analysis," *Spastics International Medical Publications,* 1968.

[11] O. Wasz-Höckert, V. Vuorenkoski, E. Valanne and K. Michelsson, "Tonspektrographische Untersuchungen des Säuglingsgeschreis," *Experientia,* vol. 18, no. 12, pp. 583-584, 1962.

[12] K. Michelsson, "Cry Analysis Of Symptomless Low Birth Weight Neonates And Of Asphixiated Newborn Infants," *Acta Paediatrica,* vol. 60, pp. 9-45, 1971.

CITED LITERATURE (continued)

[13] K. Michelsson, H. Kaskinen, R. Aulanko and A. Rinne, "Sound Spetroghraphic Cry Analysis of Infants with Hyrdocephalus," *Acta Paediatrica Scandinavica,* vol. 73, pp. 65-68, 1984.

[14] K. Michelsson, J. Raes, C. J. Thodén and O. Wasz-Höckert, "Sound spectrographic cry analysis in neonatal diagnostics. An evaluative study," *Journal of Phonetics,* vol. 10, pp. 79-80, 1982.

[15] K. Michelsson and P. Sirvio, "Cry analysis in congenital hypothyroidism.Folia Phoniatrica, 1976, 26, 40–4," *Folia Phoniatrica,* vol. 26, pp. 40-47, 1976.

[16] K. Michelsson, P. Sirvio, M. Koivisto, A. Sovijarvi and O. Wasz-Höckert, "Spectrographic analysis of pain cry in neonates with cleft palate," *Biology of the Neonate,* vol. 26, p. 353–358, 1975.

[17] K. Michelsson, P. Sirvio and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," *Acta Paediatrics Scandinavica,* vol. 66, pp. 611-616, 1977.

[18] K. Michelsson, P. Sirvio and O. Wasz-Höckert, "Sound spectrographic cry analysis of infants with bacterial meningitis," *Developmental Medicine and Child Neurology,* vol. 19, pp. 309-315, 1977.

[19] K. Michelsson, N. Tuppurainen and P. Aula, "Sound spectrographic cry analysis of infants with karyotype abnormality," *Neuropediatrics,* vol. 11, pp. 365-376, 1980.

[20] J. Tenold, D. Crowell, R. Jones, T. Daniel, L. McPherson and A. Popper, "Cepstral and stationarity analyses of full-term and premature infants' cries," *The Journal of the Accoustical Society of America,* vol. 56, no. 3, p. 975, August 2005.

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

[22] H. L. Golub, *A physioacoustic model of the infant cry and its use for medical diagnosis and prognosis,* MA, USA: Ph.D. Thesis, Massachusetts Institute of Technology, 1980.

[23] B. F. Fuller, "Acoustic discrimination of three types of infant cries," *Nursing Research,* vol. 40, no. 3, pp. 336-340, June 1991.

[24] G. Rapisardia, B. Vohr, W. Cashore, M. Peuckera and B. Lester, "Assessment of infant cry variability in high-risk infants," *International Journal of Pediatric Otorhinolaryngology,* vol. 17, no. 1, pp. 19-29, February 1989.

CITED LITERATURE (continued)

[25] B. Vohr, L. Barry, G. Rapisardi, L. O'Dea, L. Brown, M. Peucker, W. Cashore and W. Oh, "Abnormal brain-stem function (brain-stem auditory evoked response) correlates with acoustic cry features in term infants with hyperbilirubinemia," *The Journal of Pediatrics,* vol. 115, no. 2, pp. 303-308, August 1989.

[26] E. Drummond, M. McBride and C. Faye Wiebe, "The Development of Mothers' Understanding of Infant Crying," *Clinical Nursing Research,* vol. 2, pp. 396-441, 1993.

[27] M. Petroni, M. Malowany, C. Johnston and B. Stevens, "Classification of infant cry vocalizations using artificial neural networks (ANNs)," *International Conference on Acoustics, Speech, and Signal Processing,* vol. 5, pp. 3475 - 3478, May 1995.

[28] Q. W. R. L. C. Xie, "Automatic Assessment of Infants' Levels-of-Distress from the Cry Signals," vol. 4, no. 4, July 1996.

[29] A. Fort and C. Manfredi, "Acoustic analysis of newborn infant cry signals," *Medical Engineering and Physics,* vol. 20, no. 6, pp. 432-442, September 1998.

[30] J. &. R. G. C. Orozco, "Detecting pathologies from infant cry applying scaled conjugate gradient neural networks," Bruges, Belgium, 2003.

[31] M. Hariharan, J. Saraswathy, R. Sindhu, W. Khairunizam and S. Yaacob, "Infant Cry Classification to Identify Asphyxia Using Time-frequency Analysis and Radial Basis Neural Networks," *Expert Syst. Appl.,* vol. 39, pp. 9515-9523, 2012.

[32] M. Hariharan, R. Sindhu and S. Yaacob, "Normal and hypoacoustic infant cry signal classification using time-frequency analysis and general regression neural network," *Computer Methods and Programs in Biomedicine,* vol. 108, no. 2, pp. 559-569, November, 2012.

[33] M. Hariharan, L. S. Chee and S. Yaacob, "Analysis of Infant Cry Through weighted linear prediction cepstral coefficients and probabilistic neural networks," *Journal of Medical Systems,* vol. 36, no. 3, 2012.

[34] K. Kuo, "Feature Extraction and Recognition of infant cries," 2010.

[35] Y. Abdulaziz and S. Ahmad, "Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients," 2010.

CITED LITERATURE (continued)

[36] O. Reyes Galaviz and C. Reyes Garcia, "Infant cry classification to identify hypo acoustics and asphyxia comparing evolutionary neural system with a neural network system," in *4th Mexican international conference on Advances in Artificial Intelligence*, Monterrey, Mexico, 2005.

[37] E. Amaro-Camargo and C. A. Reyes-García, "Applying Statistical Vectors of Acoustic Characteristics for the Automatic Classification of Infant Cry," in *Third International Conference on Intelligent Computing*, Qingdao, China, August 2007.

[38] L. R. Rabiner and S. R. Marvin, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal,* vol. 54, no. 2, pp. 297-315, 1975.

[39] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal," in *In American Society for Engineering Education (ASEE) Zone Conference Proceedings*, 2008.

[40] C. C. L. R. R. S. N. U. J.L. Flanagan, "Synthetic Voices for Computers," vol. 7, no. 10, October 1970.

[41] T. Parsons, Voice And Speech Processing, New York: McGraw Hill Book Company, 1987.

[42] R. Priemer, MATLAB for Electrical and Computer Engineering Students and Professionals: with Simulink, SciTech Publishing, Inc., July 2013.

[43] J. Makhoul, *Linear prediction: a tutorial review,* vol. 63, Proc. IEEE, 1975.

[44] L. R. R. W. Rabiner, Digital Processing Of Speech Signals, Englewood Cliffs, N.J.: Prentice-Hall, 1978.

[45] P. Papamichalis, Practical Approaches to Speech Coding, Englewood Cliffs, NJ: Prentice Hall, 1987 .

[46] L. a. J. B. Rabiner, Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Prentice Hall, 1993 .

[47] S. &. K. K. Theodoridis, Pattern Recognition, Orlando, FL: Academic Press, 2003.

[48] D. O'Shaughnessy, "Speaker Recognition," *IEEE Magazine,* October 1986.

[49] S. Furui, "Cepstral analysis technique for automatic speaker verification," vol. 29, no. 2, April 1981.

[50] X. B. Li and D. O'Shaughnessy, "Modified Linear Discriminant Analysis for Speech Recognition," *Canadian Conference on Electrical and Computer Engineering,* pp. 1598 - 1601, April 2007.

CITED LITERATURE (continued)

[51] S. Basu, "A Linked-HMM Model For Robust Voicing And Speech Detection," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, April 2003.

[52] S. V. Kartalopoulos, Understanding Neural Networks and Fuzzy Logic: Basic Concepts and Applications, Piscataway, NJ.: Wiley-IEEE Press, 1997.

[53] J. T. Burger, "A Basic Introduction to Neural Networks," University of Wisconsin, [Online]. Available: http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html. [Accessed July 2014].

[54] F. Benini, C. C. Johnston, D. Faucher and J. Aranda, "Topical anesthesia during circumcision in newborn infants," *Journal of the American Medical Association,* vol. 270, no. 7, pp. 850-853, August 1993.

[55] C. C. Johnston and D. O'Shaughnessy, "Acoustical attributes of infant pain cries: discriminating features," in *Preceedings of the Vth World Congress on Pain*, Hamburg, Germany, August 1988.

APPENDIX

The two main MATLAB programs used in this project are given below.

```
*****************FEATURE VECTOR EXTRACTION*************
function [Feature_Vector] = AIO_preprocess(x,Fs);

    L=length(x);
    NFFT = 2^nextpow2(L); % Next power of 2 from length of y
    X = fft(x,NFFT)/L;
    f = Fs/2*linspace(0,1,NFFT/2+1);

lp_filter = fir1(40, 0.75,'low');
LPFx=conv(x,lp_filter,'same');

    L2=length(LPFx);
    NFFT = 2^nextpow2(L2); % Next power of 2 from length of y
    LPFX = fft(LPFx,NFFT)/L2;
    LPFf = Fs/2*linspace(0,1,NFFT/2+1);

Pre_emph = filter([1 -.95],[1],LPFx);   %removes DC components and flattens
spectrum to reduce effects of glottal waveform
L1=length(Pre_emph);
y=Pre_emph;

%Hamming window to compute short term energy.
    %we want a 32msec frame, with 10msec step size.

 Frame_size=.032;
 Frame_shift=.010;
 window_length=Frame_size*Fs;
 step_size=Frame_shift*Fs;
 sum1=0;
 energy=0;

%Hamming Window
w=window(@hamming,window_length);

%Compute Short-Term Energy
jj=1;
    for i=1:(floor(L1/step_size)-ceil(window_length/step_size))
        for j=(((i-1)*step_size)+1):(((i-1)*step_size)+window_length)
            y(j)=y(j)*w(jj);
            jj=jj+1;
            yy=y(j)*y(j);
            sum1=sum1+yy;
        end
        energy(i)=sum1;
        sum1=0;
        jj=1;
    end
    w=0;
    STE=energy';
```

APPENDIX (continued)

```
%Compute zero crossing rate
w=window(@hamming,window_length);

sum11=0; zcr=0;
kk=1;
    for m=1:(floor(L1/step_size)-ceil(window_length/step_size))
        y(((m-1)*step_size)+1)=y(((m-1)*step_size)+1)*w(kk);
        kk=kk+1;
        for k=(((m-1)*step_size)+2):(((m-1)*step_size)+window_length)
            y(k)=y(k)*w(kk);
            kk=kk+1;
            yy=y(k)*y(k-1);
           if(yy<0)
                sum11=sum11+1;
           end
        end
        zcr(m)=sum11/(2*window_length);
        sum11=0;
        kk=1;
    end
    w=0;
    ZCR=zcr';


% ST Autocorrelation
sum111=0; AC=zeros(size(STE),256); Au=zeros(size(STE),11);
CEPSTRUM=zeros(size(STE),11);
for u=1:(floor(L1/step_size)-ceil(window_length/step_size))
    v=1; zz=0;
    for z=(((u-1)*step_size)+1):(((u-1)*step_size)+window_length)
        zz(v)=y(z);
        v=v+1;
    end
    autocorrelation=zeros(1,256);
    for q=0:(length(zz)-1)
        sum111=0;
        for h=1:(length(zz)-q)
            s=zz(h)*zz(h+q);
            sum111=sum111+s;
        end
        autocorrelation(1+q)=sum111;    %autocorrelation matrix for window u
    end

    AC(u,:)=autocorrelation;

    %Computation of LPC based on "Levinson-Durbin" & error
    r=0;
    r=AC(u,:);
    p=10;
    [a,e] = levinson(r,p);
    Au(u,:)=a;
```

APPENDIX (continued)

```
    %Cepstral coeffiecients(first 10 LPC derived cepstral coefficients)
    alpha=a;
    cep=0;
        for d=1:(p+1)
            cep(d)=alpha(d);
            for t=1:d-1
                cep(d)=cep(d)+(t/d)*cep(t)*alpha(d-t);
            end
        end

     CEPSTRUM(u,:)=cep;
end

Feature_Vector=[STE ZCR CEPSTRUM];

(2)****************AUTORUN**********************

clear all;
tic

for i=1:288

   x=['MNRS10\MNRS10_5min\MNRS10_',num2str(i),'.wav'];
   [x, Fs]=wavread(x);
   [Feature_Vector] = AIO_preprocess(x,Fs);
   Feature_Vector(:,3)=[];
   Feature_Vector=Feature_Vector';
   load net_MNRS743gda;
   outcome = sim(net,Feature_Vector);
   y=['MNRS743219\RMNRS743219_10gda\RMNRS10nopostprc_',num2str(i)];
   save(y,'outcome')
   SmoothMA=smooth(outcome,200); %moving average with a span of 200 (2
seconds)
   index = find(SmoothMA>(.45)); %locate all indices greater than 0.45
   numberOfElements = length(index); %sum the number of frames that were
greater than 0.45

   Smooth_cry_sum_sec=numberOfElements/99; %number of seconds that were
classified as cry

smoothTime=['MNRS743219\RMNRS743219_10gda\RMNRS10_SMOOTHcrysum\RMNRS10_SMOOTH
crysum_',num2str(i)];
   save(smoothTime,'Smooth_cry_sum_sec')
   clear all;
   i=i+1;

end
**********************************************
```

VITA

NAME:	Bahare Naimipour
EDUCATION:	B.S., Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, Illinois, 2006 M.S., Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, Illinois, 2014
HONORS:	IEEE Outstanding Contribution Award, April 2013 MSU Graduate College Fellowship**,** January 2007-August 2007
PROFESSIONAL MEMBERSHIP:	IEEE Signal Processing and Education Society
PUBLICATIONS:	X. Liu, Y. Deng, B. Naimipour, D. Conger, Z. Zeng, and L. Udpa, "Evaluation of EC and GMR Sensor Response via Numerical Modeling," *Proceedings of ASNT Fall Conference and Quality Testing Show*, 2007 B. Naimipour, S. Handrick, J. Furst, D. Raicu, "Binning Strategies Evaluation for Tissue Classification in Computed Tomography Images," *SPIE Medical Imaging Conference*, 2006
WORK EXPERIENCE:	Project Engineer, Sound Answers Inc., 2013-present

UNIVERSITY OF ILLINOIS
AT CHICAGO

Office for the Protection of Research Subjects (OPRS)
Office of the Vice Chancellor for Research (MC 672)
203 Administrative Office Building
1737 West Polk Street
Chicago, Illinois 60612-7227

Approval Notice
Initial Review – Expedited Review

January 22, 2008

Tracy Magee, PhD, RN, CPNP
Nursing
845 S. Damen
M/C 802
Chicago, IL 60612
Phone: (312) 996-2193 / Fax: (312) 996-8871

RE: **Protocol # 2008-0024**
"Development of an Objective Measure of Infant Crying: A Validation Study"

Dear Dr. Magee:

Members of Institutional Review Board (IRB) #2 reviewed and approved your research protocol under expedited review procedures [45 CFR 46.110(b)(1)] on January 15, 2008. You may now begin your research.

Your research meets the requirements for review under expedited review procedures [45 CFR 46.110] Category: 6

(6) Collection of data from voice, video, digital, or image recordings made for research purposes.

Please note the following information about your approved research protocol:

Please note that proposed key research personnel, *Roland Priemer*, could not be added at this time as his investigator training certification expires on 26 January 2008 and he will not be eligible to engage in research protocols submitted to the UIC Institutional Review Board (IRB). All investigators and key research personnel involved in human subjects research must complete a minimum of two hours of continuing education in human subjects protection every two years. For more information regarding UIC's continuing education requirements, please visit the OPRS website at http://tigger.uic.edu/depts/ovcr/research/protocolreview/irb/education/continuing.shtml. Additionally, IRB staff will be happy to help identify continuing education options that best suit the needs of individual key research personnel.

Protocol Approval Period: January 15, 2008 - January 13, 2009
Approved Subject Enrollment #: 40

Additional Determinations for Research Involving Minors: The Board determined that this research satisfies 45CFR46.404, research not involving greater than minimal risk. Therefore, in accordance with 45CFR46.408, the IRB determined that only one parent's/legal guardian's permission/signature is needed.

Performance Site: UIC

Sponsor: None

Research Protocol:

a) Development of an Objective Measure of Infant Crying: A Validation Study

Recruitment Materials:

a) Flyer, Who: Infants 1 month to 4 months old who cry excessively; Version 1; 01/03/2008
b) Flyer, Who: Any infant 1 month to 4 months old; Version 1; 01/03/2008

Assent:

a) A waiver of child assent has been granted under 45 CFR 46.116(d) and 45 CFR 46.408(a) for infants too young to provide assent

Parental Permission:

a) R:\RO3 resubmission\IRB\Informed Consent; Version 1; 01/03/2008

Please note the Review History of this submission:

Receipt Date	Submission Type	Review Process	Review Date	Review Action
01/08/2008	Initial Review	Expedited	01/15/2008	Approved

Please remember to:

→ **Use only the IRB-approved and stamped consent document(s) enclosed with this letter when enrolling new subjects.**

→ Use your **research protocol number** (2008-0024) on any documents or correspondence with the IRB concerning your research protocol.

→ Review and comply with all requirements of the,
"**UIC Investigator Responsibilities, Protection of Human Research Subjects**"

Please note that the UIC IRB has the right to ask further questions, seek additional information, or monitor the conduct of your research and the consent process.

Please be aware that if the scope of work in the grant/project changes, the protocol must be amended and approved by the UIC IRB before the initiation of the change.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact the OPRS office at (312) 996-1711 or me at (312) 996-2014. Please send any correspondence about this protocol to OPRS at 203 AOB, M/C 672.

Sincerely,

Sandra Costello
IRB Coordinator, IRB # 2
Office for the Protection of Research Subjects

Enclosures:

1. **UIC Investigator Responsibilities, Protection of Human Research Subjects**
2. **Parental Permission:**
   a) R:\RO3 resubmission\IRB\Informed Consent; Version 1; 01/03/2008
3. **Recruiting Materials:**
   a) Who: Infants 1 month to 4 months old who cry excessively; Version 1; 01/03/2008
   b) Who: Any infant 1 month to 4 months old; Version 1; 01/03/2008

cc: Joan Shaver, Nursing, M/C 802

Title: Synthetic voices for computers
Author: Flanagan, J.L.; Coker, C.; Rabiner, L.; Schafer, R.W.; Umeda, N.
Publication: IEEE Spectrum
Publisher: IEEE
Date: Oct. 1970

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.