INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
SRI CITY, CHITTOOR

A Study on Infant Vocalizations towards Classification of Para–linguistic Sounds

by

Shivam Sharma

under the supervision of

Dr. Viswanath Pulabaigari

A thesis submitted in partial fulfillment for the
degree of M.S. by Research

in the
Department of Computer Science and Engineering

in

June 2019

Certificate

This is to certify that the work contained in this thesis, titled "A Study on Infant Vocalizations towards Classification of Para–linguistic Sounds" by Shivam Sharma, has been carried out under my supervision and is not submitted elsewhere for a degree.

Dr. Viswanath Pulabaigari Date
Associate Professor
Indian Institute of Information Technology, Sri City, Chittoor. A. P.

Declaration of Authorship

I, SHIVAM SHARMA, declare that this thesis titled, 'A Study on Infant Vocalizations towards Classification of Para–linguistic Sounds' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this Institute.
- Wherever I have consulted the published work of others, that is always clearly attributed by giving a proper reference.
- Wherever I have quoted from the work of others, the source is always acknowledged with a proper reference. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: _____

Date: _____

"There is no greater agony than bearing an untold story inside you."

-Maya Angelou

Abstract

Crying is known to be a crucial vocal communication medium through which an infant conveys emotions and experiences. The acoustic signal carries sufficient clues about the cause of infant's crying, the perception of which is easy for a mother or a doctor, but it is challenging to automate such recognition through a machine. Differentiation in the cry is largely a resultant of distinct vocal fold characteristics such as the tract which mostly consists of soft tissues. This results in a highly varying sound production mechanism that causes acoustic characteristics to vary significantly. Automatic identification of cry-cause factors has vast applications in assistive healthcare, and therefore can help in taking timely remedial measures in critical situations.

For studying and extracting acoustic patterns from infant cry signals, that characterize them with respect to various pre-determined causes, an infant cry signals database (IIIT-S ICSD2) is collected from paediatrics clinic. Contributions of the thesis are given below.

- **Acoustic analysis of cry sounds** is performed on the cry data to characterize cry sounds with respect to the following two objectives.
  - Finding discriminating characteristics that represent the excitation source and cry production system behaviour, for different cause factors.
  - Identifying different sound sub-types constituting an infant's cry sound, based on sound quality and acoustic parameters.

  The acoustic parameters or low-level descriptors (LLD) considered in the analysis are (i) f_0 contour, (ii) sub-band spectral energy and (iii) formant frequencies. Preliminary analysis is performed using spectrograms and by visual inspection of temporal progression of these LLDs. We then derive suitable statistical parameters for these low-level descriptors that could further help characterize cry sounds at the categorical level. The characterization is considered at different levels of granularity. At a broader level categorization, we consider a set of crying-causes, the implications of which are either severe or non-severe in nature. For example, pain and stranger's anxiety represent definite forms of severity. Whereas, discomfort and environmental change do not. At finer level categorization, these categories are considered individually. Observations and subsequent experiments show that categories that are severe in nature have several peculiarities in common, which are easily distinguishable from the ones occurring in non-severe cries. But at finer level of categorical granularity, the key distinctions observed are primarily in terms of the changes occurring at the excitation source level, the characterization of which requires the cry sounds to be represented in a way that captures these variations in a reliable manner.

- **Infant cry cause classification** is done using a set of spectral parameters. These are defined using low-level descriptors and statistical functionals that are obtained from acoustic analysis of cry signals. The feature-set obtained is first evaluated towards classifying cry sounds between categories that represent either some form of severity or absence of it. Cepstral based features and their derivatives, are empirically observed to give good classification accuracies (CA). The classifiers used for these tasks are support vector machine and multilayer perceptron, where the performances of both are observed to be reasonably good. Same features when assessed for distinctive nature towards classification among categories at finer level, i.e., for pain and stranger's anxiety, the classification accuracy (CA) is found to be of a reduced value. An end-to-end convolutional neural network, when trained over raw cry and zero frequency filtered signals, is observed to perform classification in an optimal and efficient manner, whereas raw cry based inputs were found not to give good CA for pain vs. stranger's anxiety classification. The approach of performing classification and learning the suitable acoustic representations at the intermediate convolutional layers, in a joint manner, optimizes both these sub-tasks involved in the process. Selective component filtering of the input signals based upon the acoustic parameters, observed from the cumulative frequency responses learned by the CNNs, helps validate the significance of these parameters towards the optimal classification.
- **Baby sounds classification** is done by implementing the approach of end-to-end learning using CNNs, on *baby sounds sub-challenge* data-set [1], that has different baby sounds like canonical, non-canonical, laughing, crying and a fifth category called junk. Two different approaches are examined towards classifying different baby sounds into 5 categories given in the data-set. These are listed below.
  - A monolithic approach involving implementation of a single CNN.
  - A Modular approach, where the multi-class classification task at hand is divided into smaller sub-tasks for sub-sets of fewer categories, designed according to the signal content represented by the category being considered.

  The signal content specified for the second approach is defined in terms of the presence of either speech-like or para–linguistic information within the baby sounds. The output scores of the sub-tasks are then fused into a vector and a multilayer perceptron is trained over it towards the 5-class classification objective. Although with little improvement, the modular approach of classification is observed to give better performance as compared to the monolithic approach. This experiment also suggests the significance of the network parameters that a CNN learns when trained over an infant cry data-set. The knowledge from such a network when

transferred over to another task like classifying crying vs. laughing, is observed to provide significant improvement as compared to the implementation without it.

Encouraging insights are derived from the empirical examination of the acoustic characteristics of infant vocalizations and classification approaches evaluated on different infant sounds. These insights represent cues that could further help with tasks involving development of automated assistive health-care technologies, for infants.

Acknowledgements

I take this opportunity to express my gratitude, firstly towards my mentor and advisor Dr. Vinay Kumar Mittal, for the constant support he provided throughout the duration of my masters. He laid the foundation for me, to build a comprehensive experience in academic research that led to numerous avenues that I couldn't have possibly thought of, if not for him. Equally heartfelt is my appreciation for my academic and thesis supervisor Dr. Viswanath Pulabaigari, who not only did provide his invaluable guidance and support to me, but also made me see the importance of pursuing State-of-The-Art in research.

A special thanks to Dr. Nizam (M.B.B.S-DCH) for giving the required permission to collect cry sounds from the inpatients, Dr. Bhavya (M.D.-Ped.) and Dr. Venkat (M.D.-Ped.) for providing their expertise and the nursing staff of Pranaam Hospital, Madinaguda, Hyderabad. I am also thankful to the parents of the infants for providing their support and consent towards collection of the cry audio samples needed for the study.

Few thankful feelings are reserved for Dr. Vasudev Varma, who has always beamed with positivity, hope and passion. I thank him for his valuable guidance and support at crucial times. A sincere thanks to Dr. Mathew Magimai Doss, who not only provided his excellent guidance, but also *worked by my side*, in taking the ideas and contributory work for this thesis ahead. I would also like to thank Dr. Suryakanth V. Gangashetty who has been supportive all along. This brings me to Ravi Shankar Prasad and Pavan Kumar Dubagunta, who have been a guide, friend and a colleague in their own ways, at times when I was in dire need of all three. I would also like to acknowledge the companionship of Pruthvi R. Myakala and Rajasree Nalumachu throughout the progress of associated projects.

Thanks to Korian Sir, who has been a go-to friend of mine since the beginning. A special thanks to Chhavi for making my time here in Sri City totally worth it. A big shout-out to the entire IIIT Sri City community, including faculties, friends, colleagues, academics/admin staff and fellow students, association with whom seemed pretty natural, for providing an environment of learning, growth, friendship, sharing and happiness. Many thanks to Ina and Rittick, for being great friends. I would now like to conclude by acknowledging love and support of my lovely sisters Divya and Pragyaa, who have always been there to help me preserve my sanity.

At last, but actually first and always, I humbly thank my parents for being so sweet, loving and patient to see me grow in my own ways.

Contents

List of Figures

List of Tables

Abbreviations

IIIT-S	Indian Institute of Information Technology - Sri City
ICSD	Infant Cry Signals Database
LLD	Low-Level Descriptor
DSP	Digital Signal Processing
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
STFT	Short-Time Fourier Transform
ACF	Auto-Correlation Function
VAD	Voice Activity Detection
SSE	Sub-band Spectral Energy
SF	Spectral Feature-set
std. dev.	Standard Deviation
MFCC	Mel Frequency Cepstral Coefficients
LPA	Linear Prediction Analysis
LPR	Linear Prediction Residual
ZFF	Zero Frequency Filtering
GCI	Glottal Closure Instant
SVM	Support Vector Machine
CNN	Convolutional Neural Network
C-M L	Convolution and Max-pooling Layer
FCL	Fully Connected Layer
OL	Output Layer

Symbols

f_0	Instantaneous fundamental frequency	Hz
F_i	i^{th} formant frequency	Hz
μ_{f_0}	Average f_0 for each case (file)	Hz
σ_{f_0}	Standard deviation of f_0	Hz
σN_{f_0}	Normalized std. dev. of f_0, $\sigma N_{f_0} = \sigma_{f_0}/\mu_{f_0}$	Hz
μ_{dur}	Average cry duration for each case (file)	sec
S_e	Sub-band spectral energy	–
M	Mel frequency cepstral coefficients	–
σ_M	Standard deviation of M	–
ΔM	First order derivative of M	–
$\sigma_{\Delta M}$	Standard deviation of ΔM	–

Dedicated to my parents

Chapter 1

Introduction

The earliest attempt by an infant to produce speech sound can be attributed to the laryngeal articulatory mechanism that leads to a range of qualities, reflecting phonetic possibilities and stricture types. It is observed that during the first year, the infant masters the control of articulatory detailing. Articulation control in the pharynx appears to be a necessary prior step towards expanding the articulation mannerism control to oral vocal tract tools. Therefore, the acoustic cues induced by the sound produced with the help of pharyngeal articulation, along with the alternative excitation mechanisms, are also crucial towards a systematic study of infant's sound production system. The motivation being that the baby sounds can exhibit high degree of variability, since vocal apparatus as well as muscular and motor control during the production of sounds are still developing [2], hence making the task of analyzing infant vocalization challenging. Speech analysis methods have been mainly developed on adult human speech, and they typically decompose speech signal into glottal closure based primary excitation and oral vocal tract components called as system characteristics. Thus, conventional speech analysis methods can have limitations [3]. This raises concerns about the utility of the conventional speech processing techniques for analyzing and processing acoustic characteristics of a system, which is still in nascent stage. In-order to address this issue and provide solutions that can help derive critical information from the acoustic cues in infant vocalization, it is important to first consider key differences between the sound production mechanisms of an infant and an adult.

1.1 Do infants vocalize the same way as adults?

Studies have shown how an infant acquires the control over basic speech production capacity [2]. It is observed that the speech begins in the pharynx and the phonation is

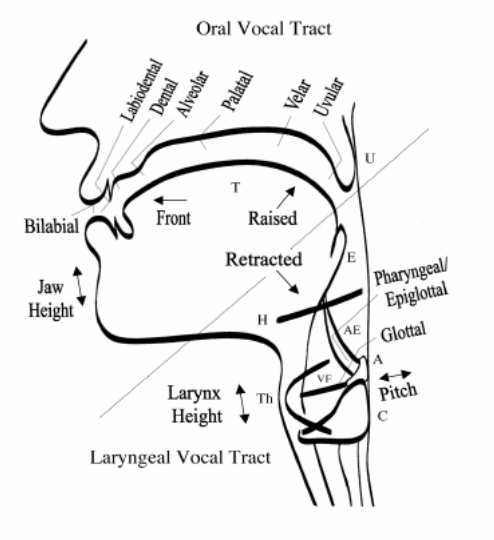


FIGURE 1.1: Depiction of the reconceptualization of the standard articulatory model into two primary articulatory domains in the laryngeal articulatory model.

assisted by the constrictions in the larynx. The laryngeal articulation includes glottal, supra-glottal and pharyngeal/epiglottal mechanisms and three vocal levels of excitation source: vocal folds, ventricular folds and aryepiglottic folds. So, alternative source vibrations and pharyngeal resonances influence the acoustics significantly. Surprisingly, it is also observed that laryngeal articulation plays a key role in distinguishing linguistic phonetic meaning in a significantly large number of languages of the world [4–7]. This could be the main cause of cultural influence that infants develop within their vocalization activities in their infancy.

Figure 1.1 shows the re-conceptualization of the standard articulatory model into two primary domains in the laryngeal articulatory model [8]. The laryngeal section of the vocal tract has a series of valves that have their own functionality and structure. The figure builds on the idea of laryngeal vocal tract and the involvement of aryepiglottic folds in the production of different sounds. Irrespective of the differences in the vocal tract shapes of an infant's sound production system from that of an adult, the consideration of the effects of laryngeal articulation and secondary source of excitation is crucial towards the acoustic analysis and the study of infant sounds like cry, babbling, canonical sounds, etc.

1.2 Motivation

In infants, the vocal cords being in primal stage and articulators not having fully evolved (the problem is further complemented by the initiation of exercising control over preliminary articulation activity in the laryngeal zone) demand utmost importance to the

dynamic intricacies of infant sound production, in performing tasks that involve acoustic analysis of infant vocalizations. Due to significant differences in the speech/sound production mechanisms for infants from that of adults, setting of the required parameters for acoustic analysis seems a challenging research task and that motivated the thesis.

1.3 Objectives and Scope of the Thesis

The objectives of the work documented as part of the thesis are listed below.

1. Deriving acoustic patterns from infant cry sounds, towards their characterization with respect to different pre-determined crying-causes. Also, correlating peculiar trends with different factors that affect cry acoustic analysis and finding whether they are indicative of additional para–linguistic information like infant's health or factors like age and psychological state.
2. Identifying the set of spectral features and alternative source representations, that optimally represent the distinctive characteristics required for the task of crying-cause classification.
3. Comparative examination of different approaches, towards obtaining the one that is empirically observed to perform the classification task in an efficient manner.
4. Validation of the spectral cues and approaches established empirically, towards their utility as part of systematic evaluations concerned with infant vocalizations.

To achieve the above objectives, the thesis did the analysis of the acoustic characteristics of infant cry sounds where the techniques used for source analysis are auto-correlation, LP Residual, sub-band spectral analysis and zero frequency filtering. Classifiers like support vector machine (SVM), multilayer perceptron and convolutional neural network (CNN) are used, for the tasks involving classification of infant sounds. Further, the utility of acoustic cues learned by CNN based experiments is validated by selective frequency filtering using digital resonators, followed by retraining of the CNN.

1.4 Thesis Organization

The work documented in the current manuscript attempts to examine the acoustic characteristics of different baby sounds. We begin our investigation by characterizing infant cry sounds by analyzing the excitation source and system characteristics, for different

crying causes and acknowledge the challenges observed in doing so. Then we move on-to examining various approaches, towards classifying not only infant cries, but also different baby sounds, into pre-determined categories. The organization of the entire work documented in this manuscript along with a briefing of each chapter is given below.

- The introduction, which is the current chapter, introduces the thesis along with its objectives and scope.
- A review of the work documented as part of the literature, is done in chapter 2.
- The primary data-set on infant cry sounds, IIIT-S ICSD2 is described in detail, in chapter 3.
- Features explored and techniques evaluated as part of the current study, are mentioned in chapter 4.
- In chapter 5, cry sounds are characterized for different causes, at both levels of categorical granularity, by examining excitation source and system characteristics.
- The challenges faced during the cry cause analysis are acknowledged and a descriptive investigation is done about different cry sounds sub-types, in chapter 6.
- Automatic classification of infant cry sounds, in different cause categories is attempted by evaluating conventional machine learning and relatively recent end-to-end learning based convolutional neural networks (CNNs). This constitutes chapter 7.
- Chapter 8 examines effectiveness of end-to-end learning approach on another, multi-class classification task involving *baby sounds sub-challenge* data-set, made available as part of INTERSPEECH 2019 ComParE challenge.
- Chapter 9 provides the summary and concludes the topic discussed in this thesis, by briefly stating possible future directions of this work.
- Publications related to the work in this thesis, along with their corresponding details are given as part of chapter 10.

1.5 Conclusion

This chapter begins with a detailed description of the speech acquisition mechanism in the early stages of anatomical development of an infant. It then proceeds further to lay

the foundation of the research problems involved in the study of infant cry vocalization. The challenge of capturing the excitation source characteristics of infant vocalization, especially during crying, is also noted herewith. Besides setting the tone of the problem at hand, by expressing the motivation, the chapter also briefly lists down the organizational structure of this thesis.

Chapter 2

Literature Review

The advent of the research on para–linguistic speech analysis is traced back to the middle of the 20th century, notably through the statements given by Crystal, that defined the phenomena of para–linguistic in speech as "vocal factors involved in paralanguage" [9]. Although, with no formal connection to the linguistics, the potential of the meaningful information contained within the acoustics of an infant cry, has already started convincing medical practitioners, parents, caregivers, etc., about its diagnostic importance.

Infant cry is a quintessential bonding mechanism, triggering the environment towards alerting and caring for the needs of the child. The cry signal quite effectively induces sympathetic nervous system activation among the people around them, and paediatricians, as well as parents have often acknowledged it to be quite informative about the health condition of the child [10]. Similarly, a laugh can indicate the well-being of the child. The cry sound is usually modulated to emphasize different patterns and characteristics reflecting a distinction of the causing factors. It is also an established medium to diagnose hidden pathological issues in infants and thereafter can be used in the treatments. Studies have highlighted a notable difference in the signal characteristics for infants with different neurological conditions. The cry signal has inspiratory and expiratory phases, of variable duration, where the latter is accompanied with rhythmic patterns of sounds. C. Manfredi et al. [11] reported that infant cry is characterized by high f_0 values which keeps changing abruptly. As infants grow, they learn to produce speech sounds through interaction with the elders. Delays in acquiring these abilities can lead to speech and language-related disorders. Thus, the development of automatic methods to analyze baby sounds is of considerable interest.

Tools like short-time spectrograms and cross-correlograms have been used towards majority of the infant cry acoustic analysis, where primary assessment of the spectro-temporal characteristics of signals have led Neustien [12] and Petroni et al. [13] to derive

various insights based on temporal progression and inter-segmental cross-correlation analysis. Also, a series of experiments have been conducted by Cohen [14] and Lavner et al. [15], in which they evaluated features like pitch, formants and MFCCs, using popular classification techniques like SVM and KNN, including the neural network based classifiers like feed-forward and convolutional neural networks.

Cry signals are generally analyzed on the basis of the parameters which are usually popular with the speech systems for adults. Optimal estimation of the spectral characteristics for infant cry signals is a challenge owing to its rapidly changing nature. A significant contribution has been made through academic challenges [16, 17], targeted at the development of the techniques, features and corpus collection for the tasks like, para–linguistic research, emotion recognition from speech, speaker trait identification, etc. Excitation source characteristics of emotional speech sounds such as shout was studied by V. Mittal [18] by using modified zero frequency filtering (modZFF). The same author [19] characterized para–linguistic sounds such as laughter using linear prediction (LP) analysis and dominant frequencies. The techniques were extended [20] towards the examination of aperiodicity for expressive voices such as *Noh* voice. Although, these studies help in analysis of para–linguistic sounds for adult speech, consequentially it has caused dearth of accumulated understanding, relevant techniques and useful resources for the applications that do not involve evolved speech like characteristics, like for analysis of infant sounds in the early stages of speech development. Nonetheless, it is customary to have the knowledge of the earlier attempts towards closely related tasks.

2.1 Studies on Infant Vocalizations

Previous studies by A. Fort et al. [21, 22] have performed acoustic analysis of infant cry signals to derive parameters such as fundamental frequency (f_0) and the vocal tract formant frequencies (F_i), based on parametric and non-parametric techniques. Parametric methods are based on linear prediction based derivation of the system response, where estimation of an optimal filter order and fundamental frequency in the residual is a challenging task. Non–parametric methods are based on liftering in cepstral domain; with high and low time liftering being performed to estimate f_0 and F_i. f_0 analysis has been done by Petroni et al. [13], Cohen [14] and V. Mittal [23, 24] by implementing Welch's method, autocorrelation, FFT and modZFF. Due to high absolute values and a high degree of variability of these parameters, estimation of an optimal length to lifter is critical. Classifiers based on Gaussian mixture modeling (GMMs) using Mel frequency cepstral coefficients (MFCCs) are derived by H. F. Alaie [25] to classify normal cry signals and the pathological cry signals. Another method by Kheddache [26] models

the MFCCs along with acoustic features such as f_0 glide and F_i dysregulation using neural networks, for early diagnosis of infant pathologies. A recent study by Abbas [27] attempts to segment the expiratory and inspiratory phases in infant cries using hidden Markov models (HMMs) trained on MFCCs. The segmentation process is aimed to aid the systems designed for the automatic detection of infant pathologies. Vocalization during inspiratory and expiratory phase in infant cries has been used by S. Grau [28] and R. Orlikoff [29] to derive cues to study their health condition. A set of 15 spectral and temporal features based on signal energy, zero crossing rate, spectral centroid, roll–off, formant locations, LPCCs and MFCCs, is utilized by Chang et al. [30] for classification of causes of infant cry using SVM classifiers.

2.2 Studies on Infant Cry Phonation

Besides the conventional approaches involving acoustic analysis using the traditional spectral features towards categorical association, it is the characterization of cries with respect to the *dysphonation* within these cries, that signify serious medical conditions. Dysphonation is popularly described by Hirschberg [31] as any deviation in the fundamental parameters like pitch, timbre, intensity and noise, from the normal behaviour. Dysphonation can be considered to be a significant indicator of valuable information about the psycho-physiological state of an infant. Different characteristics of cries were measured by Kheddache [32], with notable observations being for regions having phonations with $f_0 < 750$ Hz, and hyper-phonation > 1000 Hz. Whereas dysphonation containing noise or aperiodic sound and unvoiced sounds are observed to be the ones with $f_0 = 0$ (i.e., no vocal fold vibration).

Studies by Proytcheva [33] and Kheddache [34] have established that abnormal or violent crying can have potential cues towards detecting pathologies that are either inherently present since antenatal stage or the ones that develop after birth. Cecchini [35] considered the infant cry categorization into either *anguish, anger* and *care-seeking*. This is done with the help of feed-backs from 20 females and found dysphonation as the best indicator of anger or anguish and absence of hyper-phonation as cue for care-seeking. Pathological dysphonation was studied by Hirschberg [31] in detail, wherein 20 different types of dysphonation sub-types like *hollow, shrill* and *mew* are characterized using pitch derivatives, glide, shift-break and bitonality. Surprisingly, the crying conditions involving fussing or acoustic dysphonation of any type are not just restricted to an infant's health. Fujiwara et al. [36] studied the crying and analyzed the association between the fussing and unsoothable crying to the daily-caregivers frustration. A similar task involving feed-back based cross-validation was carried out by Zeskind [37], in which the

ratings were given by 20 abusive and 20 reference (non-abusive) parents to the crying cases studied, for being *sick, arousing* and *urgent*.

A concern regarding the developments in this field is the shortage of publicly available data-sets. The handful of data readily available is not meant for categorical studies. Another direct challenge posed is about the disparateness of the categories being studied. Studies in the past have been done for a variety of causes ranging from pathologies like asthma to disorders like asphyxia, upper respiratory tract infection (URTI), septicemia, malnutrition, congenital heart disease, etc. by Chittora [38], Wahid [39], and Chandralingam [40]. This diversifies not only the utility of the characteristic feature-set, but also the understanding about the approaches suitable for a class specific study, which although streamlines cause specific discoveries, but ultimately leads to difficulty in comparison with the State-of-the-Art in the area and hence the development remains largely localized. All these limitations are imposed by the unavailability of infant cry data-set in public domain. It is this lack of understanding and common frameworks with respect to the resources and techniques that the observations made and a fundamental approach adopted in this work, towards infant cry acoustic analysis and cause classification, is motivated from.

2.3 Conclusion

This chapter details through a series of developments that took place as part of the literature, for the research towards understanding the meaning of infant cries through machine. Conclusively, features f_0, MFCC, LPCC and spectrogram based classifications haven't yielded an optimal performance yet, in a manner in which the characterization can be generalized. Moreover, studies also highlight the significance of phonation modalities present within a cry sound that are often correlated with different pathological states. There are few directions, in which there is a clear requirement of research and development, wherein highly varying characteristics of infant vocalizations are truly captured. An attempt has been made to address a fragment of the issues concerned, in the current thesis.

Chapter 3

Infant Cry Signals Database (IIIT-S ICSD2)

First attempt from the lab at IIIT-S, towards investigating infant cry acoustics was done in [41, 42]. As part of these works, the spectrograms of different infant cries for same infants but for different causes were examined. The observations were cross-validated using the parameter plotting and comparing, which elucidated the distinctive characterizing capability of f_0 contour for different categories. The data-set used was *IIIT-S ICSD*, which was initially collected from a paediatrics clinic, under the supervision of paediatricians and nurses. Besides for the objective of qualitative examination, it was the requirement of a larger data-set, with proper ground-truth information that solicited *IIIT-S ICSD2*, first introduced in [43].

Details on IIIT-S ICSD can be referred from [41, 42]. The details about the ComParE Baby Sounds Sub-challenge (BSS) data-set, made available as part of INTERSPEECH 2019, is also used for specific studies in this work and can be referred in detail from the challenge paper [1]. Only IIIT-S ICSD2 has been focused upon as part of the data-set description in this manuscript, since it forms the basis of a majority of the study conducted in this work. Apart from this, the required details on baby sounds data-set *BSS* are mentioned wherever necessary and as per the need.

3.1 Building the Data-set

The IIIT-S ICSD2 data-set is also collected from the same clinic, as for IIIT-S ICSD. The Zoom H4n recorder was used to collect the data, at a sampling rate of 48 kHz. The recorder has a built-in X/Y configuration of stereo mics to suppress the ambient noise

TABLE 3.1: Cause, age and gender-wise cry count distribution in IIIT-S ICSD2 data-set

Cry-causes	*Age in months (Male)*			*Age in months (Female)*			*# cry bouts*	*Duration (min.)*
	< 3	$3-18$	> 18	< 3	$3-18$	> 18		
Pain	113	212	80	158	341	74	978	33
Anxt	0	58	40	25	130	34	287	18
Ailt	0	19	0	0	7	0	26	1
Disc	5	0	6	0	47	15	73	3
Envr	0	46	106	0	0	21	173	6
Hung	0	0	0	64	0	0	64	2
Emot	0	5	0	0	3	0	8	1
Overlap	0	75	14	0	11	0	100	3
Total	118	415	246	247	539	144	1709	67

and provide a clean recording at a close-speaking distance. The data-set is collected over a total of 104 (50 male and 54 female) infants, mostly aging between 6–15 months, with a maximum age of 6 years. Table 3.1 gives the details of the data collected against factors like cry causes, gender and age of infants. For additional analysis of the cry sounds, the cause categories are also studied as part of a broader and a finer level categorization. The broader level categorization is defined for severe and non-severe crying causes. These categories form a super-set of the finer level-categories, depending upon the level of severity they represent. Whereas, the finer level categorization is directly based upon the causes (ground-truth) for which the cry data has been collected.

Infrequency for the causes such as hunger/thirst, ailment and emotional need was observed during the recording sessions, and therefore there are fewer samples for these categories, which are not considered as part of the study. The labeling of cry signals according to different causes was done as per the observations of paediatricians, based on on-going and historical medical conditions, parent's inputs and infant's health status. The data contains background noise, mostly by other patients and nurses. Pain either due to vaccination or pathological condition, along with stranger's anxiety are major causes of cries which are severe in nature. Discomfort and environmental change are other less severe cause categories which lead to cries. Therefore, the database is also grouped into severe vs. non-severe causes of infant cry, for specific studies as part of this work. Cry sound sub-types observed from the data-set are *growl, shrill (mostly strained, with few softer variants too), squeal* and *moan*. The proposed categorization of cry-causes at different levels of categorical granularity is shown schematically in Fig.

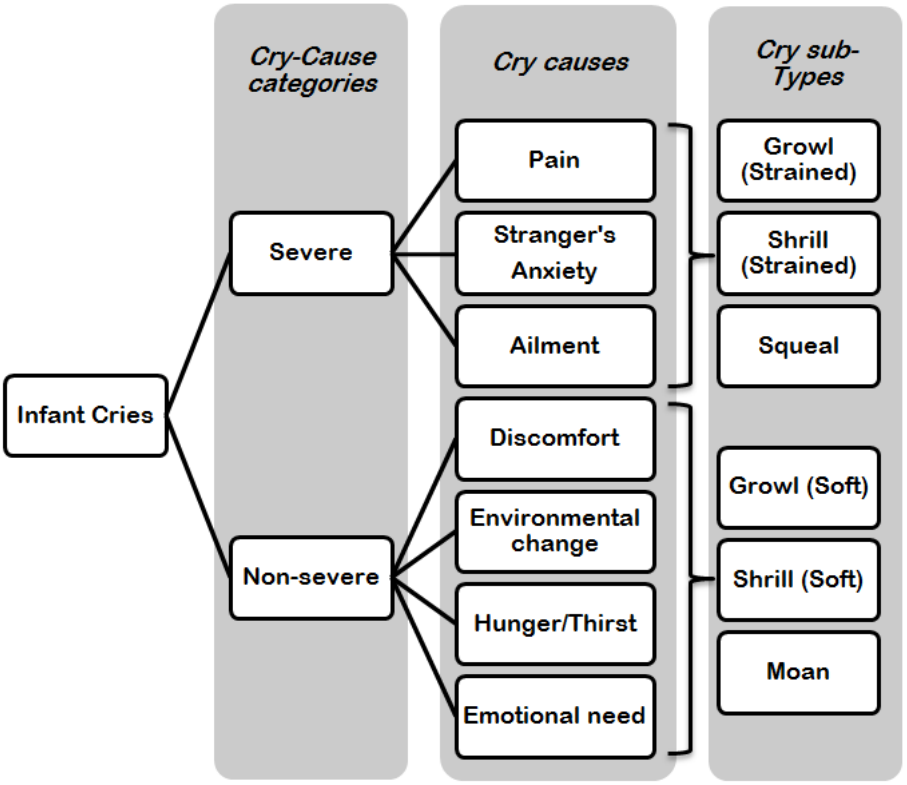


FIGURE 3.1: Schematic diagram of cry-causes, and sub-type categorization at different levels.

3.1. As a reference to such a categorization, the distinction between different classes based on their causes has been proposed in literature [44].

3.2 Conclusion

The primary data-set on infant cry sounds used in this study, IIIT-S ICSD2 has been described in detail in this chapter. In the interest of collaborative research and development in this area, the IIIT-S ICSD2 data-set is expected to be released in the public domain shortly. Although on request basis, IIIT-S ICSD has already been provided in the public domain for research and non-commercial purposes.

Chapter 4

Features Explored and Techniques Evaluated

Infant cry data-set IIIT-S ICSD2 is first acoustically analyzed using several fundamental digital signal processing techniques. These techniques are used to examine the progression of excitation source characteristics and the variation in articulatory configuration with time. Characterizing features identified in this process are further evaluated towards their discriminative ability with respect to different pre-determined cause categories, using different classification approaches. This chapter is dedicated to briefly introducing the features explored and techniques considered towards acoustic analysis, and subsequently for classification based tasks performed as part of this work.

4.1 Features explored and associated DSP techniques evaluated

Features considered in this study are briefly described in the following sub-section.

4.1.1 Features explored

4.1.1.1 Short-time Spectrogram

This is obtained by processing the segments of the cry signals in the frequency domain, which is given by,

$$X(\tau, \omega) = \sum_{n=-\infty}^{n=\infty} x[n]w[n-m]e^{-j\omega n} \tag{4.1}$$

with $x[n]$ being the signal and $w[n]$ being the window function [45]. This gives spectrogram and energy, used for preliminary acoustic analysis of a cry signal.

4.1.1.2 Harmonics

The periodicity of the signal is clearly seen during short-time analysis, which manifests in the form of fundamental frequency and harmonics, as narrow peaks at equally spaced frequencies in the short-time spectrum [46].

4.1.1.3 Instantaneous Fundamental Frequency (f_0) Contour

Instantaneous fundamental frequency or f_0 based contour is a key parameter for analyzing excitation source characteristics [13]. In our primary analysis, the contours from the f_0 estimates are plotted using a State-of-the-Art algorithm YIN [47], which served as a reference here. These f_0 estimates are colored yellow (basic), green (good) and blue (best). These are then compared with the contour derived using the methods below:

1. *Autocorrelation:* The f_0 contour is computed using autocorrelation with window length 30 ms and a shift of 10 ms.
2. *Linear Prediction (LP) residual:* As a secondary validation LP residual is used with autocorrelation, having window length 30 ms shift 10 ms, LP order 14 and down-sampling of 10 kHz.

4.1.1.4 Sub-band Spectral Energy

Sub-band spectral energy is derived by performing filter-band analysis. The ratios of sub-band spectral energies (ϵ), for the bands under consideration are computed to get correlated patterns for these *sub-bands*. It is computed by performing the dot product of power spectrum ($S[k]$) and the filter banks (f_b) and is given as [48],

$$\begin{aligned} \epsilon &= S[k].f_b \\ where, S[k] &: NX(\frac{N_{fft}}{2}+1), and \\ f_b &: N_{filt}X(\frac{N_{fft}}{2}+1) \end{aligned} \tag{4.2}$$

with, N, N_{fft} and N_{filt} are # of frames, FFT bin size and # of filter-banks respectively.

4.1.1.5 Formant Frequencies

Formant frequencies derived using the LP spectrum are analyzed for examining the changes in the vocal tract filter, i.e., the system characteristics [49]. First 5 formants $F_1 - F_5$ are examined. A frame size of 30 ms with a frame shift of 10 ms and LP order 50 are used for the LP analysis of the signals having 48 kHz sampling rate.

4.1.1.6 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is one of the most commonly used features for front-ends in the speech recognition systems. The feature vectors are extracted from the spectra of windowed cry signal. For a Mel scale cepstrum of order 'p', the feature vector is obtained by considering the first p DCT coefficients [50]. Generally, p is set as 13, and the extracted coefficients are denoted as M. A total of 52 coefficients, including MFCCs M, delta MFCCs ΔM, and their standard-deviations, σ_M and $\sigma_{\Delta M}$ respectively, with 13 coefficients each are computed.

4.1.1.7 Zero Frequency Filtered (ZFF) Signals

The ZFF signal is obtained by filtering the speech through an ideal digital resonator centered at 0 Hz [51]. The equivalent transfer function of the filter is given by,

$$H(z) = \frac{1}{1 - z^{-1}}. \tag{4.3}$$

The equivalent time domain operation is an integrator given by,

$$y[n] = y[n-1] + s[n], \tag{4.4}$$

where $s[n]$ is the differenced and mean removed input signal. Application of the filtering results in a polynomial type growth trend in the output signal. The filtered signal therefore has to be trend removed across duration of the approximate pitch period to obtain the ZFF signal.

Features explained above are computed by implementing the following associated digital speech/signal processing techniques.

4.1.2 DSP Techniques used for feature extraction

4.1.2.1 Autocorrelation

Autocorrelation provides a measure of the similarity given by

$$r_x(m) = E[x(n)x(n+m)] \tag{4.5}$$

between the waveforms of the time functions, i.e., the source signal $x(n)$ and the delayed version of itself $x(n+m)$ [52]. f_0 contour is derived using this technique.

4.1.2.2 LP Analysis

The prediction of the current sample as a linear combination of past p samples forms the basis of LP analysis, where p is the order of prediction [49]. As $A(z)$ is the reciprocal of $H(z)$, the LP residual is obtained by inverse filtering of the speech

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{4.6}$$

First 5 formants are derived from the LP spectrum.

4.1.2.3 Cepstral Analysis

The log of the power spectrum is considered to be the linear sum of the information representing the speech sound production characteristics and source excitation characteristics. These can be retrieved by processing in the quefrency domain. The resulting quantities are called as a cepstrum, the low-frequency components of which give insight into the production characteristics and high-frequency components are excitation source characteristics [50].

4.1.2.4 Filter-bank Spectral Analysis

Filter-bank analysis is used for the spectral analysis in different frequency bands [53]. Sub-band spectral energy represents the spectral energy (ϵ) with respect to the critical frequency based bands under consideration.

4.1.2.5 Zero Frequency Filtering

The essence of zero-frequency filtering lies in computing the output of the cascade of two zero-frequency resonators [54], which is equivalent to four times successive integration of x[n], the signal. The expression for this is given below,

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{4.7}$$

Where, a1 = -2, and a2 = 1. This operation as mentioned previously, repeated twice.

The discontinuities within a zero frequency filtered signal are obtained by subtracting 10 ms (ideally equivalent to a pitch period) moving average at each sample from this signal. These discontinuities within the source excitation can be obtained from the instants of positive zero crossings, indicating the glottal closure instants (GCIs). These instants can be anchored upon to estimate the instantaneous fundamental frequency (f_0).

4.1.2.6 Digital Resonator Design

A digital resonator [55] is a special two pole band pass filter with the pair of complex conjugate poles located near the unit circle. The angular position of the pole determines the resonant frequency of the filter. In the design of a digital resonator with a resonant peak at or near $\omega = \omega_0$, we select the complex-conjugate poles at,

$$p_{1,2} = re^{\pm j\omega_0}, \quad 0 < r < 1 \tag{4.8}$$

4.2 Machine Learning Techniques used for Classification

Towards the objective of performing classification of different infant vocalizations, into pre-determined causes/factors, following algorithms have been used.

4.2.1 Support Vector Machines (SVM)

The support-vector network implements the following idea: it maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori. In this space, a linear decision surface is constructed with special properties that ensure high generalization ability of the network [56].

4.2.2 Multilayer Perceptron (MLP)

Data enters at the inputs and passes through the network, layer by layer, until it arrives at the outputs. During normal operation, that is when it acts as a classifier, there is no feedback between layers. This is why they are also called feed-forward neural networks [57].

4.2.3 Convolutional Neural Network (CNN)

CNN is designed to automatically and adaptively learn spatial hierarchies of features through back-propagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers [58].

4.3 Conclusion

This chapter gives a detailed list of features and techniques evaluated in this work. The objective is to provide a brief introduction about the fundamental idea behind them and their usage in this work. The reader is requested to redirect to the correspondingly cited resource for a deeper understanding of any individual component. Also, specific parameterization, as and wherever done with respect to any feature or a technique is clearly stated, along with the necessary context provided, to maintain the continuity.

Chapter 5

Infant Cry Cause Analysis

The cry audio data is collected for multiple categories. Cries from different categories are perceptually distinct, and the distinction pattern is perceived in terms of factors like signal intensity, energy, pitch variation, cry duration, etc. In-order to perform qualitative analysis, short-time spectrograms are observed for various cries. Based upon the patterns of harmonics, formants, fundamental frequency and spectral energy observed from the spectrograms, distinctive characteristics are found to exist for the cries from different causes. These characteristics are then quantitatively analyzed by deriving suitable low-level descriptors and statistical functionals that can effectively capture the acoustic characteristics of the cry signals for different causes.

Towards these objectives, the chapter begins by stating steps involved during the pre-processing stage of the cry database. This is followed by performing preliminary analysis using short-time spectrograms. Based upon the observations made, cry sounds are characterized using basic statistical measures. Acoustic parameters like f_0, sub-band spectral energy and formant frequency are analyzed in further sections, for different types of cry sounds.

5.1 Pre-processing of Infant Cry Audio Data

During the audio data collection stage, live speech recognition systems are often required to be complemented by possible hints of words and phrases to improve the accuracy for specific words and phrases [59]. Whereas, there is no prior requirement of this type towards the collection of cry data. Although, the recording do need to emphasize on regions having most prominent presence of cry bouts, as against any imminent ambient noise present in the background.

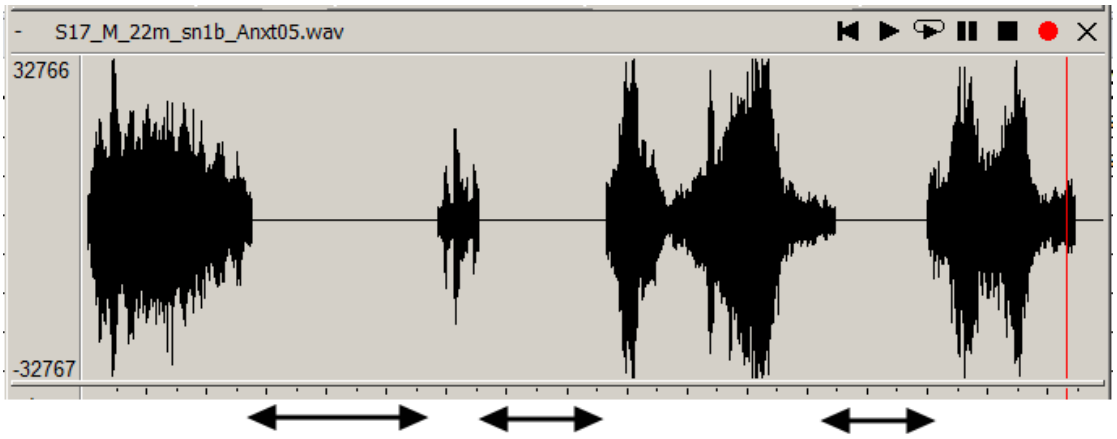


FIGURE 5.1: Waveform with *silenced (arrow-marked)* and *cry* audio parts.

Keeping such differences in the perspective, several simple considerations are made during pre-processing of the raw cry audio data which is recorded from the clinic. These are enumerated below,

1. The more the number of data files, the bigger and better the training data that is required for machine learning tasks. Therefore the recording sessions are divided into excerpts appropriately. Cry sounds with larger gaps in between two continuous sub-sessions, are also split as per the requirement.

2. The intermediate babble noise is transformed into the silence regions as shown in Fig. 5.1.

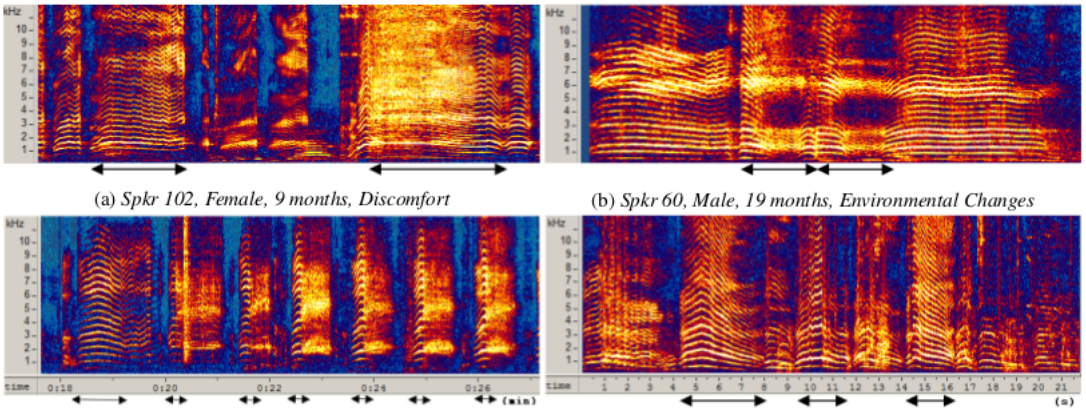
3. The intermediate pause information (presence and duration) is preserved.

4. Even if the infant crying is extremely inaudible in some regions, with no significant sound being perceived, these are silence-transformed too. It helps preserve intra-cry duration required for the study.

5. There is no point in processing the background disturbances for the purpose of analyzing cry acoustic information, especially when the analysis is happening at fundamental level. Therefore, such regions are ignored.

6. The raw data collected is preserved to facilitate for the studies examining robustness in the presence of the background noise.

The above-mentioned points briefly summarize the manner in which the entire data-set is processed.



(a) *Spkr 102, Female, 9 months, Discomfort*

(b) *Spkr 60, Male, 19 months, Environmental Changes*

(c) *Spkr 7, Female, 18 months, Pain*

(d) *Spkr 67, Male, 12 months, Stranger's Anxiety*

FIGURE 5.2: Illustration of differences in the contours of f_0 and its Harmonics observed in the spectra of infant cries for 4 different categories. (a) Discomfort cry (monotonously flat contours), (b) Cry due to environmental changes (flat contour with lesser fluctuations), (c) Pain cry (short inverted cup-shaped contours) and (d) Cry due to strangers anxiety (prolonged inverted cup-shaped contours) [Please observe the changes in the arrow marked regions].

5.2 Preliminary Analysis using Short-time Spectrogram

At first, spectrograms are observed for characteristic spectral behaviour. These are 4096 point FFT'd, short-time spectrogram with Hamming as analysis window type, and pre-emphasis factor of 0.97, in the popular speech analysis tool called Wavesurfer [21]. The cry-cause categories examined are: *discomfort, environmental change, pain* and *stranger's anxiety*. In each category 15, 15, 16 and 20 cry samples, respectively, are used for analysis.

In Fig. 5.2 (a), it can be observed that spectrogram for the discomfort based cry is mostly flat with not many variations in the pitch and harmonics. But because of some high pitched growling effects, energy distribution appears to coexist with cry outs. Environmental change category has a characteristic intense shrill effect in the sound, appearing as high-pitched harmonics. This category also, has relatively stable pitch progression, with lesser short duration pitch fluctuations. Such cries often occur with the random combination of various cry types like brisk cry outbursts, shrill sounds and occasional growling (Fig. 5.2 (b)). Pain cries have clear and intense spectral characteristics with peculiar brief inverted cup-shaped cry events (Fig. 5.2 (c)). Notably, a transition of the sound type is observed from squeaky to growl effect, while studying the cry behaviour for the infants ranging from neonatal stage through the infancy of over 9 months. The cases for strangers anxiety also have hyper-phonated (non-normal) sound types along with similarly shaped (arched) cry events but are observed to be relatively prolonged (Fig. 5.2 (d)).

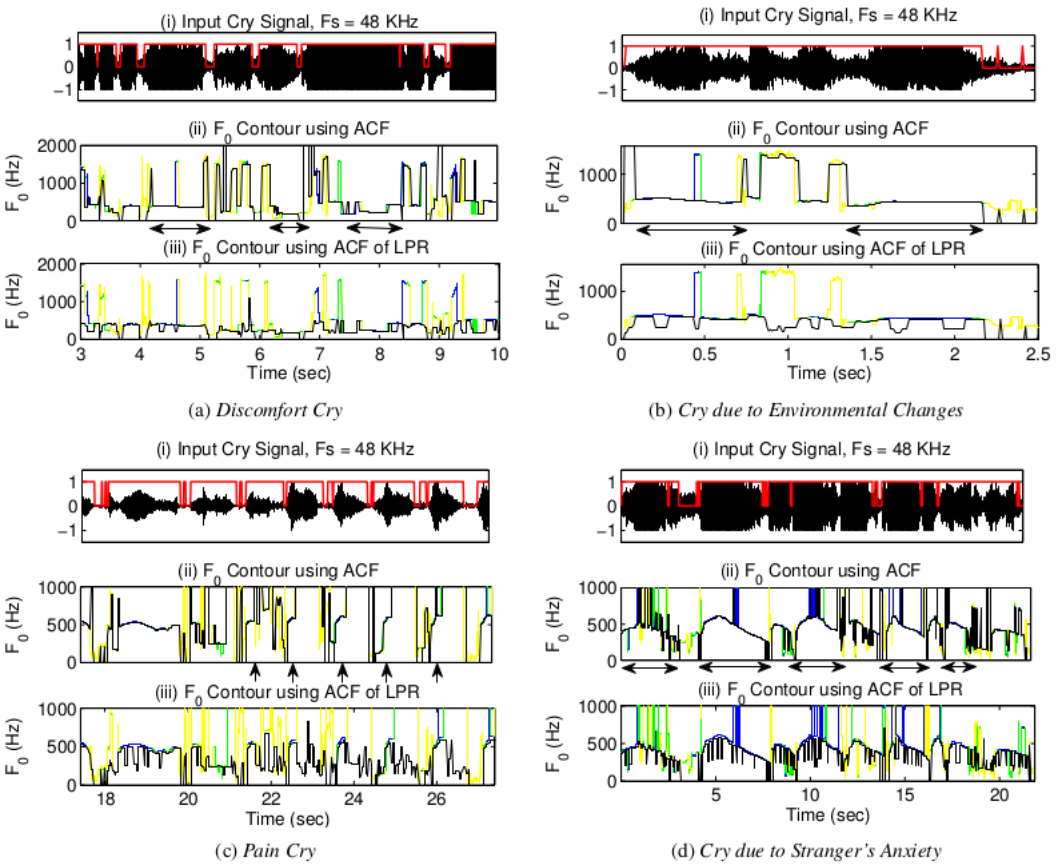


FIGURE 5.3: Illustration of differences in f_0 contours of infant cries, evaluated using auto-correlation function (ACF) and LP residual (LPR), for 4 different categories. (a) Discomfort cry (monotonously flat con- tours), (b) Cry due to environmental changes (flat contour with lesser fluctuations), (c) Pain cry (short inverted cup-shaped contours) and (d) Cry due to strangers anxiety (prolonged inverted cup-shaped contours) [Please observe changes in the arrow-marked regions in the f_0 contours obtained using the ACF (middle subplots)].

The variations in the cry acoustic signals are significantly higher, that are observed to be induced also with the growing age, showing intense changes from 9 months on-wards. Also, the presence of different sound types that make up a cry vocalization, induce significant fluctuations in the acoustic characteristics of the signal. The experiments conducted are upon the infants lying mostly between the age group of few days up-to 20 months. Also, at times the presence of different sound types can be crucial for analysis, which is why they have been duly studied in this work.

5.3 Characterization using Excitation Contour

5.3.1 Analysis using f_0 contour

Firstly, the unwanted background noises are manually removed by performing silence transformation of the noisy regions. The silence transformed cry signals are then observed after the application of a voice activity detection algorithm [60] and marked in red, as observed in Fig. 5.3 (i). This ensured that the analysis and processing is done only for the parts contributing to the prominent cry regions. The short-time analysis of the cry signal is performed by considering frame size of 30 ms and frame shift of 10 ms. Due to the dynamic nature of the vocalizations within infant cry sounds, contours for the discomfort cries induce moderate deflections, especially at the cry event onsets (Fig. 5.3 (a)). The fluctuations for the various cry sound types observed for environmental changes as shown in Fig. 5.3 (b), are observed to be sparse due to their minimal transitions. These are the 2 categories having relatively more stable (i.e. the one having less average pitch deviation) cry melody contour. Pain cry bout contour (Fig. 5.3 (c)) brings out the cyclic f_0 fluctuations, also validating the spectrogram based observations mentioned as part of section 5.2. The dysphonated cry sounds like growling, squeaking, etc., can occur for this category too, but the core cry zone will have prominent arc-shaped cry patterns that occur frequently, but briefly. With significantly intense f_0 fluctuations (in comparison with that of pain) as can be seen in Fig. 5.3 (d), stranger's anxiety cries produce relatively prolonged cry patterns.

From this analysis, peculiar acoustic and perceptual patterns have come forth as decisive towards a high level distinction between the cries for different causes. The cry signals with majority of the cry events having less contour deviation and more stability are generally due to the causes that are not that severe in nature for instance, discomfort and environmental change, whereas if there are arc shaped patterns (described above) present within the core cry regions, it might indicate severity, like for the cases due to pain or the presence of some stranger (stranger's anxiety). The observations related to excitation contour analysis are briefly enlisted as below:

- Discomfort cry has stable contour with moderately high ($\approx 500 - 1000\ Hz$) fluctuations.
- Environmental changes cry events have sparse f_0 fluctuations.
- Pain cry has brief, frequent inverted-cup shaped spectral patterns varying over $100 - 200\ Hz$.

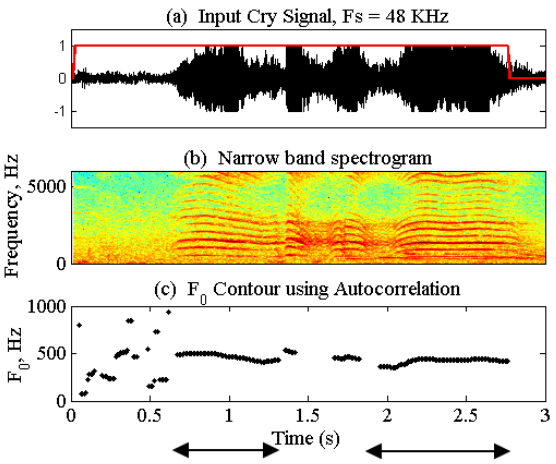


FIGURE 5.4: Acoustic features for infant cry due to *environmental change* [Spkr 60, *Male*, 19 months]. (a) signal, (b) Narrow-band spectrogram, (c) f_0 contour (using autocorrelation) [Notice flat contours in the arrow-marked regions].

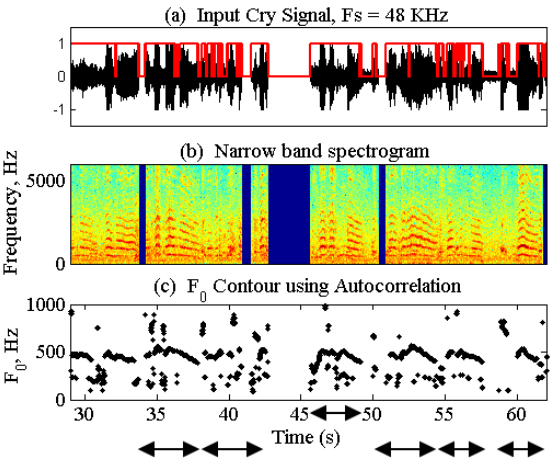


FIGURE 5.5: Acoustic features for infant cry due to *stranger's Anxiety* [Spkr 62, *Female*, 15 months]. (a) signal, (b) Narrow-band spectrogram, (c) f_0 contour (using autocorrelation) [Notice inverted cup-shaped contours in the arrow-marked regions].

- Stranger's anxiety has gradually descending (> 2 *sec*) cry events with significantly large (≈ 1500 *Hz*) f_0 fluctuations.
- Cries having less deviations in the f_0 contour, usually indicate a non-severe cause category, like discomfort or environmental change.
- Cries having more and frequent deviations in the f_0 contour indicate a severe cause category, like pain or stranger's anxiety.

There is significant development in the vocal chords and the cavity that induces lot of changes in the cry acoustics. These manifest in the form of high fluctuations in the cry melody and f_0 contour. Although, the absolute size of the infant cry production system is small as compared to that of an adult, and the resulting cry outbursts generally tend to be brief in nature, the overall patterns do show some peculiarity in different situations. It is during the non-severe crying cases, like the ones due to discomfort or environmental change, that the melody and harmonic patterns during the core cry event are relatively more stable (i.e. with less changes), whereas they are observed to be of frequent and significant changes for the cases from more severe categories like pain and stranger's anxiety, occurring in the form of brief cry outbursts due to psychological reasons. The analysis done and the patterns observed consistently reflect the characteristics of 66 samples analyzed, thus providing insightful inputs towards further exploratory work.

TABLE 5.1: Quantitative analysis using parameters to measure melody contour characteristics - (c) Average, (d) Std. Dev., (e) Normalised Std. Dev., of f_0 and (f) Average cry segment-duration for different cases of environmental change.

(a) Infant (M/F)	*(b)Age (Months)*	*(c)* μ_{f_0} (Hz)	*(d)* σ_{f_0} (Hz)	*(e)* σN_{f_0} (Hz)	*(f)* μ_{dur} (sec)
S44 (M)	9	553.8	249.8	0.45	1.1
S30 (M)	11	348.8	193.7	0.56	1.2
S30 (M)	11	303.4	149.8	0.49	0.8
S70 (M)	12	539.1	199.5	0.37	0.7
S12 (M)	18	399.8	210.3	0.53	0.9
S60 (M)	19	573.6	329.5	0.57	1.3
S83 (F)	12	503.7	269.1	0.53	1.7
S78 (F)	20	428.1	224.8	0.53	0.9
S05 (F)	36	488.2	354.0	0.73	1.4
S59 (F)	44	516.8	342.2	0.66	1.0
Average		465.5	252.3	0.54	1.1

5.3.2 Analysis of variation in f_0 contour

The input cry signals after removing the unwanted background babble noise are processed post application of the voice activity detection. The prominent cry regions are identified and processed accordingly. These cry regions are distinctly depicted in the input signals in Fig. 5.4 (a) and 5.5 (a). Spectrogram is used for observing characteristic spectral behaviour. This is a 512 point FFT'd short-time spectrogram with Hamming window, and pre-emphasis factor of 0.97. The final sub-plot analysis is for the variation in instantaneous fundamental frequency (f_0) with time.

5.3.2.1 Distinguishing environmental change cries from stranger's anxiety cries

Our earlier observations from section 5.3.1 are further consolidated in terms of the f_0 statistics in this analysis. Further comparative analysis for the categories environmental change and stranger's anxiety showed very subtle differences with respect to the same features. For stranger's anxiety, one significant difference based upon the patterns of the cry melody contour is the presence of prolonged arch-shaped cry signal segments, similar to the ones found within the pain cry cases, but relatively few. Hence, logically the category average of the std. - dev. in the f_0 contour for this category should be more than that for environmental change, but the empirical verification suggests otherwise (see Tables 5.1 and 5.2, column (d)). The reason could very well be the combined effect of

TABLE 5.2: Quantitative analysis using parameters to measure melody contour characteristics - (c) Average, (d) Std. Dev., (e) Normalised Std. Dev., of f_0 and (f) Average cry segment-duration for different cases of stranger's anxiety.

(a) Infant (M/F)	*(b)Age (Months)*	*(c)* μ_{f_0} (Hz)	*(d)* σ_{f_0} (Hz)	*(e)* σN_{f_0} (Hz)	*(f)* μ_{dur} (sec)
S26 (M)	6	322.7	187.5	0.58	1.1
S67 (M)	12	455.3	197.2	0.43	1.4
S70 (M)	12	519.2	297.4	0.57	1.8
S60 (M)	19	824.8	227.4	0.28	1.7
S79 (M)	24	417.8	160.1	0.38	1.6
S28 (F)	2	302.0	182.4	0.60	2.4
S87 (F)	6	567.7	402.6	0.71	1.0
S85 (F)	13	501.5	270.3	0.54	1.2
S24 (F)	13	305.9	157.7	0.52	1.9
S69 (F)	15	449.6	203.6	0.45	1.0
Average		466.7	228.6	0.51	1.5

different types of hyper-phonated (or non-normal) cry sounds that make up most of the core cry segments present within the cry sounds available for environmental change, along with the relative scarcity of the arch-shaped cry patterns in the case of stranger's anxiety. A comparison between these characteristics is shown for brief portion of cries from both the categories in Fig. 5.4 and 5.5, leading to statistical analysis at the cry signal segment level, observed from Table 5.1 and 5.2.

Presence of rising and falling contour shapes can be clearly seen from the spectrogram for the stranger's anxiety case considered here, as shown in Fig. 5.5, which is nearly absent for the environmental change in Fig. 5.4.

The category average of the std. - dev. of f_0 contour turns out to be lesser for stranger's anxiety as compared to that for environmental change, with very less deviation in the case-wise values of parameter σ_{f_0} as shown in Tables 5.1 and 5.2. This solicits analysis of highly localized melody contour patterns, when characterizing for these two categories.

5.3.2.2 Pitch range comparison

(1) *Environmental change and stranger's anxiety:* The values obtained from the statistical analysis for various cases, suggest some insights. The minimum average f_0 for each case/file μ_{f_0}, for the categories environmental change and stranger's anxiety, as can be observed from Tables 5.1 and 5.2, are close enough and are $\approx$ 303 Hz for both. But for the latter category, the maximum of this parameter can

TABLE 5.3: Quantitative analysis using parameters to measure melody contour characteristics - (c) Average, (d) Std. Dev., (e) Normalised Std. Dev., of f_0 and (f) Average cry segment-duration for discomfort cases.

(a) Infant (M/F)	*(b)Age (Months)*	*(c)* μ_{f_0} (Hz)	*(d)* σ_{f_0} (Hz)	*(e)* σN_{f_0} (Hz)	*(f)* μ_{dur} (sec)
S55 (M)	9 (d)	391.7	169.5	0.43	0.6
S09 (M)	11	351.5	163.3	0.46	2.9
S57 (M)	15	541.3	216.1	0.40	1.3
S41 (M)	30	591.7	237.4	0.40	0.7
S34 (F)	7	288.5	154.7	0.54	1.4
S102 (F)	9	725.6	382.3	0.53	1.0
S16 (F)	9	329.4	175.8	0.53	1.3
S09 (F)	11	805.5	586.1	0.73	1.0
S29 (F)	12	762.1	465.1	0.61	1.0
S101 (F)	24	375.7	195.0	0.52	1.3
Average		516.3	274.5	0.53	1.2

go even above 800 Hz. This could be possibly be due to the subtle stimulation at the beginning of the crying due to change in the external environment for both the cases, resulting in similar average f_0 at both category and case levels (see category averages in Table 5.1 and 5.2), but the gradual intensifying of the anxiety levels due to a stranger's presence in case of *stranger's anxiety*, tends to induce significant changes in the characteristics like cry melody, pitch etc., as explained previously with the help of Fig. 5.4 and 5.5, ultimately leading off to a high f_0 contour.

(2) *Discomfort and pain:* The case-wise dynamic range of f_0 for the discomfort turns out to be $\approx$ 40 Hz below the corresponding values for the pain category, with minimum/maximum deviations for both the categories as 154.69/586.11 Hz and 196.12/534.13 Hz respectively (Table 5.3 and 5.4). Such drastic extension of the maximum f_0 level for the discomfort category could be attributed to the different types of the hyper-phonated cry sounds present within, whereas for the pain cry cases, the increment appears to be significantly through normally phonated cry sounds. This is validated from the respective average values, which when observed with the help of the Table 5.4 for pain category, is well ahead than for the discomfort by more than 75 Hz (Table 5.3). Also, the *normalized* std. - dev. tabulated as column (e) in the Tables above, clearly has pain with the highest value.

TABLE 5.4: Quantitative analysis using parameters to measure melody contour characteristics - (c) Average, (d) Std. Dev., (e) Normalised Std. Dev., of f_0 and (f) Average cry segment-duration for different cases of pain.

(a) Infant (M/F)	*(b)Age (Months)*	*(c)* μ_{f_0} (Hz)	*(d)* σ_{f_0} (Hz)	*(e)* σN_{f_0} (Hz)	*(f)* μ_{dur} (sec)
S40 (M)	1	844.9	398.7	0.47	1.0
S25 (M)	4	324.5	534.1	1.65	1.4
S70 (M)	12	720.3	328.4	1.46	1.00
S39 (M)	13	673.3	268.4	0.40	1.00
S32 (M)	16	708.9	416.9	0.59	1.2
S73 (F)	1	627.0	248.7	0.40	0.8
S15 (F)	5	718.5	403.9	0.56	1.0
S14 (F)	7	480.0	196.1	0.41	1.3
S07 (F)	18	812.7	456.8	0.56	1.0
S38 (F)	18	670.7	384.8	0.41	1.0
Average		658.1	363.1	0.61	1.0

5.3.2.3 A perspective on cry duration

The comparison of the duration of the cry bouts, identified by the VAD does not reflect any special characteristic distinctions. Since the average cry duration for cry-cause categories environmental changes, stranger's anxiety, pain and discomfort are obtained as 1.23 *sec*, 1.02 *sec*, 1.09 *sec* and 1.50 *sec* respectively, as can be observed for the average duration values (μ_{dur}) from Tables 5.1, 5.2, 5.4 and 5.3, column (f). The sharp surge in the values for the first and last category is due to exceptional cases having the average cry duration for each case (μ_{dur}) longer than the general trend, which is either due to sustained cry signal segment or the clubbing of two consecutive cry signal segments that are too close for the voice activity detection algorithm to distinguish.

5.3.2.4 Discussion

The relative cry intensity levels for the categories under focus can also be interpreted from the perspective of the level of uneasiness the infant might be experiencing when under their influence. For the category pain, the infant would want to get rid of the unpleasantness developing within as early as possible. Resultantly, the cry sound production will be through the system, the articulatory configurations for which would regularly be in a state of conveying these intense associated emotions in various ways, the acoustic effects of which would manifest as different crying patterns within a cry signal, whereas for stranger's anxiety, the zenith of the cry intensity within the cry

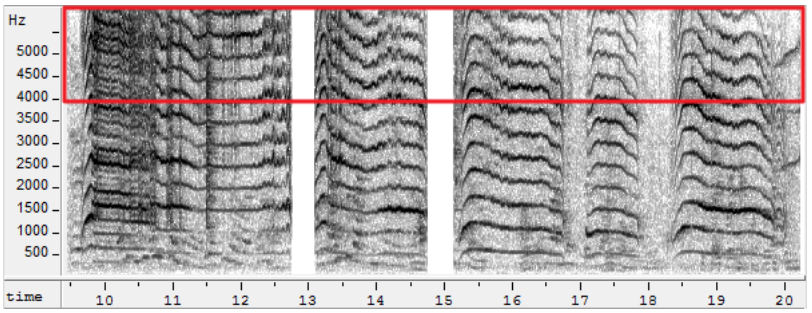


FIGURE 5.6: Spectrogram illustrating *intense* spectral characteristics with uniform distribution for *pain* cry, *Spkr # 01*.

bout, is not reached as early as for pain because of the time taken by the infant to get over the perplexity due to external stimulations accompanied along with this category. Categories like discomfort and environmental change cause an infant to stay perplexed for a significant time, thereby not generating significant acoustic variations within the melody of the cry sound.

Key observations

1. Cry segment deviation is more for pain than for discomfort by 88 Hz.
2. Cry segment deviation for environmental change is more than that for the stranger's anxiety by 23.65 Hz.
3. The minimum average f_0 for both *environmental change* and *stranger's anxiety* is $\approx$ 303 Hz, whereas its maximum reaches 800 Hz for the latter, suggesting more intensity in its cry patterns.
4. Average f_0 dynamic range for discomfort is lesser than that for pain by $\approx$ 40 Hz. The key distinction here lies in the cry melody contours for the cry bouts of these categories, which is reflected as greater category average of std. - dev. of f_0 (σ_{f_0}).
5. Average cry duration for all the categories is obtained to be 1.18 *sec*.

5.4 Characterization using Sub-band Spectral Energy

Short-time spectrograms of cry signals for categories pain, discomfort and environmental change, as shown in Fig. 5.6–5.8, depict spectral content across the spectrum for the signals. Characteristic distinction can be observed in terms of the distribution of this energy across the frequency scale. This drastic variation in the spectral energy is observed in terms of *sub-band spectral energy (SSE) ratios*. This effect is described for the categories under consideration as below,

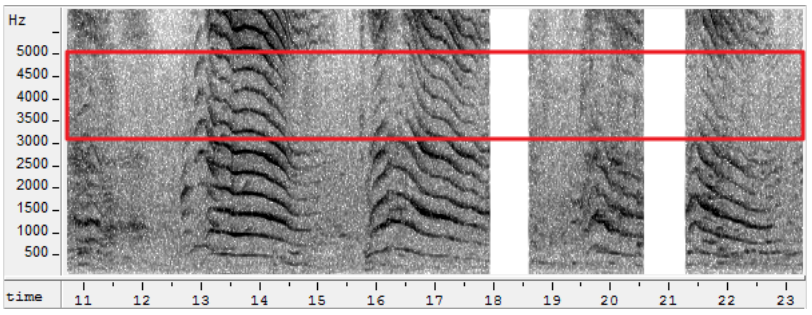


FIGURE 5.7: Spectrogram illustrating mild spectral characteristics with skewed distribution for *environmental change* cry, Spkr # 60.

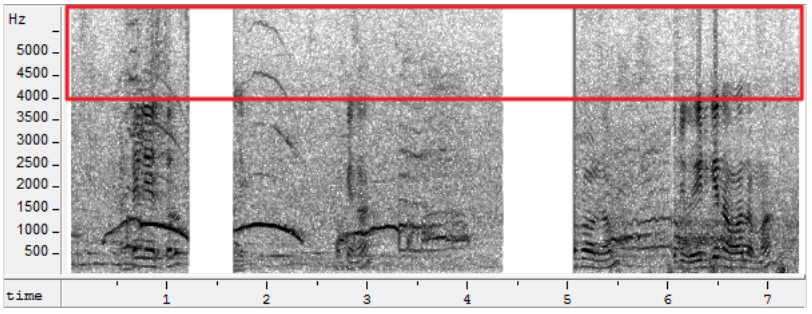


FIGURE 5.8: Spectrogram illustrating weak spectral characteristics with skewed distribution for *discomfort* cry, Spkr # 55.

5.4.1 For pain cries

As can be observed from Fig. 5.6, the spectral energy for Spkr #1 appears to be distributed *uniformly* throughout the signal spectrum, across the signal progression. This results in higher SSE values for the higher band-based ratios. Sub-band spectral energy ratios of IV w.r.t I and II sub-bands, i.e., (α_1 and α_2 are observed to be significantly high as shown in Table 5.5, columns (h) and (i).

5.4.2 For cries due to environmental change

For this category, the presence of relatively lower sub-band spectral energy is visible for higher frequency ranges as shown for Spkr #60 (Fig. 5.7). The overall spectral energy appears to be less for higher sub-bands, as compared to that for pain category. Hence, the energy ratios corresponding to IV sub-band (around 4 kHz) are higher than that for the other bands, possibly due to the presence of higher resonant frequencies for this category. The two sub-band spectral energy ratios (α_1 and α_2 are observed to be moderately low as compared to that of the pain category, as can be seen from Table 5.5.

TABLE 5.5: Average Sub-band spectral energies ((b)-(g)), for cries due to pain, discomfort and environmental change categories. The ratios of spectral energies as $\alpha_1 = \frac{E_4}{E_1}$ and $\alpha_2 = \frac{E_4}{E_2}$ are shown in (h) and (i) respectively.

(a) *Category*	*(b)* E_1	*(c)* E_2	*(d)* E_3	*(e)* E_4	*(f)* E_5	*(g)* E_6	*(h)* α_1	*(i)* α_2
Pain	.02	.11	.30	*.59*	*.44*	*.30*	*39*	*6*
Discomfort	.03	.11	.38	.19	.21	.17	6	2
Environmental-change	.02	.09	.28	.22	.14	.09	12	2

5.4.3 For discomfort cries

The sub-band spectral energies for the cries due to discomfort, like for Spkr #55, when observed using the spectrogram showed that the energy is significantly less in the higher sub-bands marked by 4, 5 and 6 kHz boundaries (Fig. 5.8). This behaviour can also be examined from the SSE ratios (α_1 and α_2, for discomfort category from Table 5.5. The values are observed to be significantly lower than those of the other 2 categories.

5.4.4 Discussion

Spectral energy in the higher frequency ranges of $4 - 6$ kHz (i.e., E_4, E_5 and E_6) give distinctive peculiarities for *severe* (pain) and *non-severe* (discomfort and environmental change) cry-causes. The observations made using sub-band spectral analysis are summarized in Table 5.5 and explained along with. Cries due to severe causes are observed to have consistently higher sub-band spectral energies, whereas non-severe cry-causes have lower sub-band spectral energies in the frequency ranges above 3000 Hz. This happens possibly due to the insufficient phonatory activity during non-severe based crying, while at the same time the infant being curious about the dynamics of the stimulus due to physiological or environmental changes. On the other hand, for the cases due to severe causes, the infant could better localize the source of suffering which for the pain induced cry could be a physical injury for instance.

5.5 Characterization using Formant Frequencies

5.5.1 Different stages in infant cry vocalizations

Just as speech is composed of different linguistic units that build the language characteristics, similarly an infant while crying tends to induce modulating effects, especially in the middle of each cry bout, mostly for severe cases. 3 different stages of a cry bout, commonly observed for a typical infant crying can be stated as below:

- Onset,
- Build-up, and
- Fading.

Cry *onset* is initiated by subtle vocal tract modulations, excited by varying amount of the lung-pressure and excitation pattern. In the second stage i.e. cry *build-up*, the energy and the formant frequencies are raised up which results in the cry intensity build-up to a higher level. This is the *core* cry region, within which lies crucial cry information such as infant's psycho-physiological state, age, cry-cause severity, etc. Any leftover air-pressure is released in the final stage, i.e., *fading* of cry.

TABLE 5.6: Average formant frequencies ($F_1 - F_5$ (in Hz)) for the 3 stages of infant cries: Stage I-'Onset' ((b)-(f)), Stage II-'Build-up' ((g)-(k)) and Stage III-'Fading' ((l)-(p)). Note that 'Build-up' formant frequencies are higher for severe cry category.

(a) Category	*Stage I - 'Onset' of Infant Cry*					*Stage II - 'Build-up' of Infant Cry*					*Stage III - 'Fading' of Infant Cry*				
	(b) F_1	(c) F_2	(d) F_3	(e) F_4	(f) F_5	(g) F_1	(h) F_2	(i) F_3	(j) F_4	(k) F_5	(l) F_1	(m) F_2	(n) F_3	(o) F_4	(p) F_5
Severe	1035	2184	3364	4839	6186	*1314*	*2355*	*3528*	*5067*	*6262*	1169	2270	3391	5023	6221
Non-severe	1050	2228	3469	5070	6112	1059	2065	3341	4912	5997	1018	2167	3434	4948	6137

TABLE 5.7: Average formant frequency differences ($\Delta F_{xy} = F_y - F_x$) (in Hz) for the 3 stages of infant cries: Stage I-'Onset' ((b)-(e)), Stage II-'Build-up' ((f)-(i)) and Stage III-'Fading' ((j)-(m)).

(a) Category	*Stage I - 'Onset' of Infant Cry*				*Stage II - 'Build-up' of Infant Cry*				*Stage III - 'Fading' of Infant Cry*			
	(b) ΔF_{12}	(c) ΔF_{23}	(d) ΔF_{34}	(e) ΔF_{45}	(f) ΔF_{12}	(g) ΔF_{23}	(h) ΔF_{34}	(i) ΔF_{45}	(j) ΔF_{12}	(k) ΔF_{23}	(l) ΔF_{34}	(m) ΔF_{45}
Severe	1149	1181	*1475*	1347	1041	1173	*1540*	1194	1101	1121	*1632*	1195
Non-severe	1177	1241	*1601*	1042	1007	1276	*1571*	1086	1150	1267	*1514*	1189

Table 5.6 and 5.7 show the formant frequencies and their differences respectively, averaged for *10* severe and *6* non-severe cases. The average formant frequency values $F_1 - F_5$ for the severe cry category in Table 5.6 are observed to be significantly high

for *Stage II* (build-up stage) as compared to those from the other 2 stages (onset and fading). Whereas no such increase is observed for non-severe cry-category (Table 5.6). It can also be observed from Table 5.7, that the difference between 4^{th} and 3^{rd} formant frequency values (*i.e.*, ΔF_{34}) is significantly greater than other ΔF values for both the severe and non-severe categories for all the stages. The ΔF values for formants $F_1 - F_4$ for non-severe cry category are consistently higher in each of the three stages of infant cry bouts. The possible reason could be the prominent usage of front vowel based articulator configurations like that for /*i*/, /*I*/, /æ/, and /ε/, for non-severe category cry sounds.

5.6 Conclusion

Investigation of acoustic characteristics of infant cry signals, for different causes is done by analyzing the excitation contour, sub-band spectral energy and variation in vocal tract system characteristics over time. The analysis sheds light on key patterns and acoustic trends present within different types of cry signals. It is with respect to these characteristics that infant cries can be differentiated based on the severity their cause-categories represent.

Investigation of production characteristics of infant cries is challenging due to minimal deviation in the vocal tract shape. Hence, distinctive spectral characteristics are not clearly perceived from formant frequency analysis for different cry-causes. However, the average formant frequencies in the *Built-up* stage highlight the significant changes in the system characteristics of *severe* category of cry-causes.

Due to the presence of significant dysphonation within cry sounds, the f_0 estimation becomes challenging. Although, a general idea about the behaviour of excitation source characteristics of cry signals is obtained from the analysis done as part of this work, but in-order to capture such characteristics reliably, a better representation parameter needs to be evaluated. A parameter that can effectively capture the effects of alternative source of excitation and a production system, the time-varying nature of which does not affect the parameter estimation as much. This is in line with the evaluations involving automated computation of cry parameters and recognition of patterns from cry signals, towards different tasks.

Chapter 6

Infant Cry Sub-type Analysis

Infants cry due to different causes, in distinctly audible sub-types of cry sounds. These cry sound sub-types are either *strained* or *softly* produced variants of sound types such as *growl, shrill, moan,* etc., that can be present in any cry sound. These different cry sound sub-types for severe and non-severe categories have already been introduced as part of this work and shown schematically in Fig. 3.1. Growl and shrill being majority of the cry sound types, are considered specifically in the current study. As a reference to a study of such sub-types of cry sounds, Hirschberg [31], as noted earlier, has reported the association of dysphonation sub-types like *hollow, shrill* and *mew* with different pathologies. An illustration of differences in spectrograms for each of the cry-sound sub-types growl and shrill cry is depicted in Fig. 6.1 and Fig. 6.2, respectively.

This chapter introduces different cry sound sub-types that are observed perceptually and empirically, in the following section. It further proceeds to build-upon the empirical observations, by characterizing these cry sound sub-types using formant frequencies and short-time magnitude spectrums.

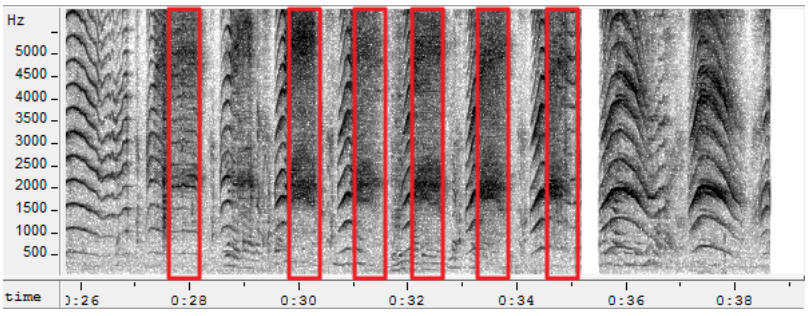


FIGURE 6.1: Spectrogram for growl cry sound [Spkr 7, Female, 18 Months].

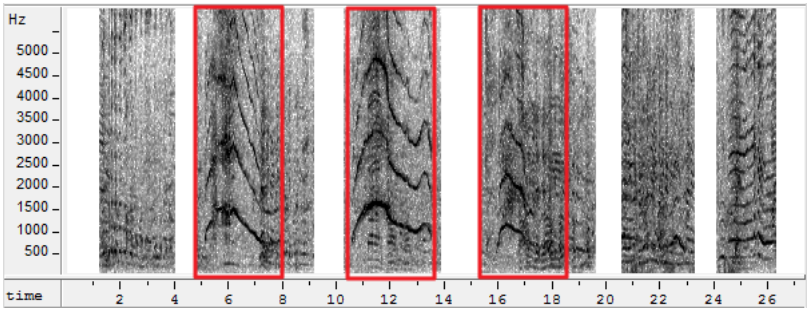


FIGURE 6.2: Spectrogram for shrill cry sound [Spkr 104, Male, 1 day].

6.1 Types of Infant Cries

Since most of the inpatients had their vaccination scheduled, one of the common reasons for the crying response of the infants was regarded as pain, by the doctor. Pain cries have been known to have peculiar acoustic characteristics; few notable ones are enlisted below:

- Homogeneous vocalization
- Briefly rising contour at the onset of a cry event
- Gradually descending contour
- Brisk expiration

A reason for such acoustic manifestations could possibly be a temporary unbearable change in the physical condition an infant's body is subjected to, which the child intends to get rid of as soon as possible. This makes the infant cry out in various kinds of ways (some enlisted above), nevertheless all in brisk succession. Empirical evidences suggest that while sometimes occurring in rhythmic patterns, the cries vary significantly with age. The rationale here could be the various tendencies that tend to build up over time within the human psyche. Some peculiar kinds of cry sounds observed from the qualitative analysis of the cry data-set IIIT-S ICSD2, along with the categories they appear with are briefly stated below,

- The clear cry voicing sound that gives distinct periodicity and well defined harmonics (mostly belonging to pain).
- Shrill is a whistle like sound. Following 2 sub-types have been observed:
  - Soft shrill cry is where we find the formant frequencies, but with lack of clear harmonics (found with the environmental changes).

TABLE 6.1: Formant frequencies ($F_1 - F_5$) for different cry sub-types, for severe and non-severe categories, respectively.

(a) Cry Sub-type	*Formant values*					*(g) Category*
	(b) F_1	(c) F_2	(d) F_3	(e) F_4	(f) F_5	
Growl (Strained)	*1506*	*2421*	*3490*	*5076*	*6187*	*Severe*
Shrill (Strained)	*1415*	2087	3481	5027	6185	
Average	*1461*	*2254*	*3485*	*5052*	*6186*	
Growl (Soft)	996	2090	3004	4432	5098	*Non-severe*
Shrill (Soft)	1079	2313	3603	4964	6134	
Average	1037	2202	3303	4698	5616	

  - Also, there is loud shrill, when the pitch and spectral energy is high.

- Cry shout is when there is a distinct loud shout mostly produced as an outcome of strong pain.
- There is a series of brief, brisk expirations (mild variants reported for the onsets of cries due to environmental change).
- Growly cry is a roar like sound having too much varying f_0 due to hyper-phonation.
- Squeaky sound that an infant produces occurs mostly in conjunction, as the following variants:
  - Normal, high pitched squeaky sound,
  - Semi-strained squeak, and
  - Fully-strained squeak.
- Moan cry is where a soft moan sound is heard. This voice can usually be heard at the ending of a cry session. It is just a weaker cry sound, having low f_0, low energy and formant frequencies similar to that of a normal cry.

In this work, cries due to a particular reason may be referred to by their category names for simplicity. Like for cry due to pain, we will have the terminology *pain cry*, unless stated otherwise. Also, cries other than normally voiced ones (like shrill, growl, squeaky, etc.) will be collectively referred to as *non-normal* cries.

Growl cry sounds have spectral energy *scattered* over a wider frequency range (Fig. 6.1), whereas shrill cry sound sub-types have it more concentrated at sparse harmonics

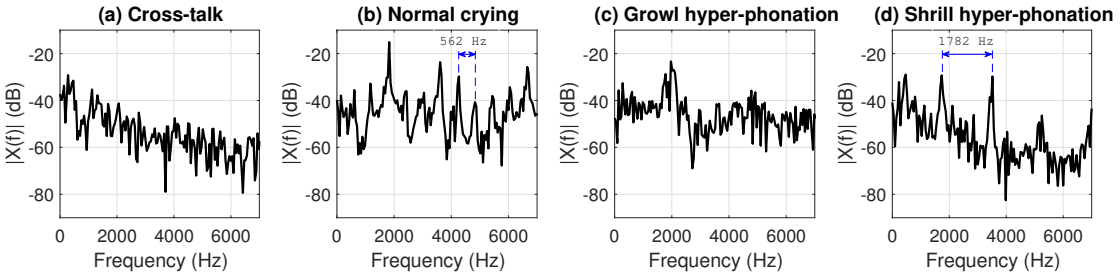


FIGURE 6.3: Short-time magnitude spectrum comparison for four different types of regions ((a)-(d)) observed in a cry signal.

(Fig. 6.2). It can be inferred from these figures, that first two formants occur in close proximity with each other for growl cry sounds (Fig. 6.1), whereas formants occur in well-spaced manner for the shrill cry sub-type (Fig. 6.2).

6.2 Characterization using Formant Frequencies

Average formant frequencies ($F_1 - F_5$) are given in Table 6.1 for different cry sub-types. F_1 is observed to be comparatively higher for strained growl sound (for severe cry category) and soft shrill (in non-severe cry category). Notably, the average $F_1 - F_5$ are highest for growl (strained) cry sub-type. Whereas, all formants $F_1 - F_5$ have higher values for shrill (soft) cry sub-type in non-severe cry category. Moreover, the formant frequency ranges observed for different cry sub-types imply the dynamic behaviour of the vocal tract configurations during cry sound production, especially during the dysphonations.

6.3 Characterization using Short-time Magnitude Spectrum

Based upon the spectral cues of the transition events for 10 different cases, some depicted in Fig. 6.1 and 6.2, analysis using frequency domain characteristics lead to some significant understanding towards characterization of different regions within a cry signal. As can be observed from Fig. 6.3, depicted are the short-time (25 ms) magnitude spectrums of *cross-talk, normal crying, growl hyper-phonation and shrill hyper-phonation* crying regions, in the respective sequence from left to right. Fig. 6.3 (a) has typical noise-like spectral characteristics, with relatively lower spectral energy along with no definite pattern of the frequency content. On the other hand, the spectrum for the normal crying, shown in Fig. 6.3 (b), as expected has definite harmonics with frequency peaks equally spaced by 562 Hz. Also, it has relatively higher average spectral energy.

Notably, the spectrum also exhibits the structure for the resonant frequencies, indicating the production characteristics of an infant instead of any other system. Fig. 6.3 (c) shows the spectrum of the *growling* region in a cry. This is marked by more spectral energy, structure depicting the resonant frequencies and *non-definite* patterns of harmonics. Whereas, *shrill* hyper-phonation is the sub-type with well-defined harmonic structure and characteristic *high-frequency* excitation. As can be seen from Fig. 6.3 (d), the harmonics are corresponding to 1782 Hz f_0. Such distinctive spectral behaviour of different cry regions can be insightful towards performing distinctive characterization, of cry sub-types, hence of different cry types.

6.4 Conclusion

The analysis of cry signals elucidated dynamic nature of the cry production characteristics representing excitation source and system. The vocalization exhibits highly varying trends in the spectral energy, f_0 contours, signal energy, etc., during the production of various crying sounds sub-types. Most common sub-types of cry sound observed during the analysis are growl, shrill and squeak. Also observed along with are the soft/strained variants of each of these.

The average formant frequencies $F_1 - F_5$ are observed to be higher for severe category with both growl and shrill being strained variants. Short-time spectrums are also observed to divulge prominent patterns for background cross-talks, normal crying and strained/soft variants of growl and shrill. The distribution observed is in terms of spectral energy, harmonicity, formant structure, etc. Wherein, for sounds like normal crying and shrill, prominent periodicity within spectral peaks is observed.

It is not just the peculiarities observed within the cry signal characterization of these sound sub-types, but also the correlation of their intensity with the growing age that renders their analysis crucial towards examining infant cries.

Chapter 7

Broad and Fine Classification of Infant Cries using CNNs

Infant cries are vocalized signals where the vibrating vocal fold characteristics embed essential voice quality related to different factors. The proposed approach attempts to classify the infant cry signals within broad and fine classes based on their causes. The data-set used for the current study is the same IIIT-S ICSD2, curated further towards evaluating machine learning based classification systems. We examine the cry signal characteristics for various causes at different level of categorical granularity. As an example of the perceptual observation, the cries for causes in severe classes are observed to exhibit longer duration and higher signal energy. A classification within these broad classes is attempted based on feature–set derived from the spectral analysis of cry signals. A feature set derived from the fundamental frequency contour, sub-band spectral energy and MFCC values along with their gradients is derived as parameters to train a support vector machine (SVM). The proposed features are tested individually, and as a set, for their ability towards the classification of the two classes. A comparison in the classification performance of the proposed parameter set is made with a MLP system.

A further classification is attempted for finer causes of infant cry in the severe class. The spectral features are used over a SVM to classify between the cry signals caused by pain and stranger's anxiety classes. The results motivate the employment of a classifier with the ability to capture intricate characteristics of the signal. A CNN based architecture is therefore used for classification using the raw signal. In a CNN based end–to–end classification, learning each step is done in a simultaneous manner, keeping all other steps and the final task in account. End–to–end architectures have recently been utilized for several applications related to fields of automatic speech recognition, gender detection,

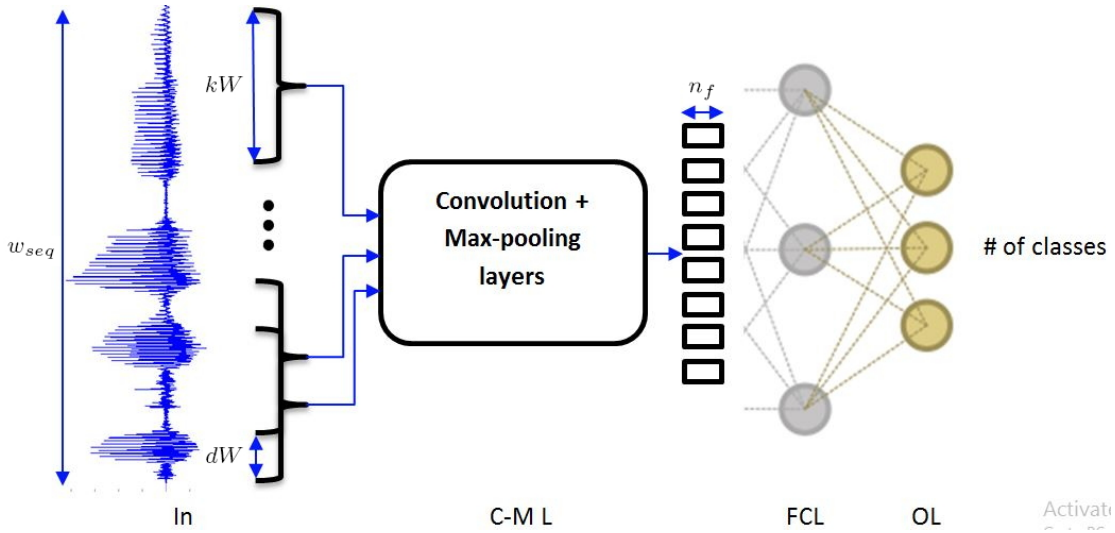


FIGURE 7.1: Illustration of first convolution layer processing and network structure [In: input; C-M L: convolution, max-pooling layer; FCL: fully connected layer; OL: output layer].

children speech recognition and speech pathology detection [61–64]. These systems are efficient towards learning an optimal windowing and spectral processing along with the features and the classifier for a specific task. An improvement of the classification is achieved using the source representation of the cry signal obtained using the zero frequency filtering (ZFF) based method.

7.1 CNN-based Raw Waveform Modeling

We adopted the CNN-based raw waveform modeling approach that was first developed for speech recognition, and later scaled to other tasks such as, speaker verification, gender recognition and depression detection. In this approach, the network architecture consists of N convolution layers (with max-pooling and ReLU) followed by one hidden layer based multilayer perceptron (MLP). During training, both the CNN and MLP parameters are trained by minimizing a cost based on cross entropy in an end-to-end manner. During testing, the output class probabilities are averaged over the utterance to make the final decision.

Figure 7.1 illustrates the first convolution layer processing with n_f filters. It takes as input signal of length w_{seq} samples and processes it with a frame size (kernel width) and frame shift (stride) of kW and dW respectively.

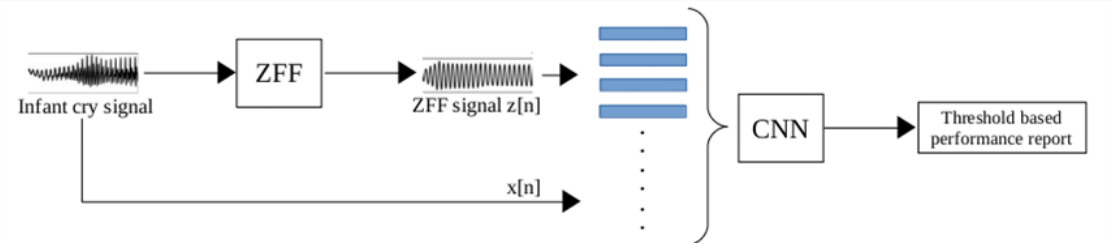


FIGURE 7.2: Illustration of implementation with different input types.

7.1.1 CNN input

We investigate two types of input: (i) raw signal and (ii) zero frequency filtered (ZFF) signal. The approach is also depicted in the Fig. 7.2. The latter input type is motivated from the fact that in speech there exists an excitation component. In baby sounds, modeling these information could be beneficial, as infant cry is characterized by high f_0 values which keeps changing abruptly [11].

7.2 Experimental Setup

7.2.1 Classification in Broad Classes

The parameters derived from spectrum consist of mean of f_0 (μ_{f_0}), standard deviation (σ) of f_0 (σ_{f_0}), sub–band spectral energies (S_e), MFCCs (M), σ in MFCCs (σ_M), differenced MFCC (ΔM), and σ in ΔM ($\sigma_{\Delta M}$). These parameters are derived over a duration of 30 ms with a shift of 10 ms. The S_e values are obtained in a band of 1 kHz in the range of 0–6 kHz. The parameters obtained from severe and non–severe classes are used to train a SVM using Gaussian kernels. A combined performance of all these parameters together is also examined using the SVM, and is further compared with a single layer MLP. The performance is also compared with end–to–end CNN architecture trained over raw signals and ZFF. The CNN uses the input segment $s_t^c = \{s_{t-c} \dots s_t \dots s_{t+c}\}$ at time t, along with a context spanning around $2c$ frame, to learn parameters. An initial learning rate of 0.1 is used and the network is trained over a minimum of 40 epochs. Stochastic gradient descent (SGD) technique is used for learning weights, with a momentum factor of 0.5. Implementation of the architecture uses Keras libraries with Tensorflow back–end, with 4 convolution layers. The training uses a kernel width (kW) of 16 along with a context, to result in an overall dimension corresponding to 250 ms, at a shift of 10 ms.

7.2.2 Classification in pain vs. anxiety causes of infant cry

A set of MFCC and its gradient ΔM, along with the σ_M and $\sigma_{\Delta M}$ values, obtained across a window length of 30 ms for cry signals corresponding to pain and stranger's anxiety classes, are used to train a SVM with Gaussian kernels. The classification results are also compared with the results obtained over MLP with a hidden layer, using the same parameters. To improve on the classification results, an end–to–end architecture is trained over raw cry and ZFF signals. The ZFF method involves filtering the speech through an ideal digital resonator centered at 0 Hz [51]. Application of the filtering results in a polynomial type growth trend in the output signal. The filtered signal therefore has to be trend removed across duration of the approximate pitch period to obtain the ZFF signal. An ideal trend removal process requires a good estimate of pitch of the signal. The ZFF signal is a good representation of the speech source and carries information related to voice quality, emotions, and other para–linguistic elements [65, 66]. Representation of cry signals based on parameters derived from the spectrum is not adequate due to rapidly changing fundamental frequency and other production characteristics in the signal. The broad classes as well as the finer classes classification is attempted using a 4 layer CNN architecture with kW=16 and kW=300.

TABLE 7.1: Classification between severe and non–severe classes using spectral parameter set SF.

Features	*Precision*	*Recall*	F-1
μ_{f_0} and σ_{f_0}	0.7	0.7	0.7
S_e	0.6	0.68	0.63
M	0.98	0.98	0.98
σ_M	0.97	0.97	0.97
ΔM	0.93	0.93	0.93
$\sigma_{\Delta M}$	0.98	0.98	0.98
Average	0.86	0.87	0.86

7.3 Results

7.3.1 Broad and fine class classification

The spectral features set $SF = \{\mu_{f_0}, \sigma_{f_0}, S_e, M, \sigma_M, \Delta M, \sigma_{\Delta M}\}$ is obtained for the given data-set. A SVM is trained over each of these parameters and the results are given in the Table 7.1. The results show that MFCC and its derivative, along with their mean and variance values give a good classification score. Source features as the μ_{f_0} and σ_{f_0}, or the sub-band spectral energy ratio do not lead to a good classification performance.

TABLE 7.2: Classification between severe and non–severe classes of infant cry causes.

Features	*Architecture*	*Precision*	*Recall*	*F*-1
SF	SVM	0.97	0.97	0.97
SF	MLP	0.96	0.96	0.96
Cry signal	CNN (4-L)	0.99	0.99	0.99
ZFF signal	CNN (4-L)	0.99	0.99	0.99

TABLE 7.3: Classification between pain vs. anxiety causes of infant cry.

Input	*Architecture*	*Precision*	*Recall*	*F*-1
MFCC	SVM	0.94	0.95	0.94
$M, \sigma_M, \Delta M, \sigma_{\Delta M}$	MLP	0.78	0.78	0.78
Cry signal	CNN (kW = 16)	0.77	0.76	0.75
	CNN (kW = 300)	0.76	0.76	0.76
ZFF signal	CNN (kW = 16)	0.99	0.99	0.99
	CNN (kW = 300)	0.98	0.98	0.98

The set *SF* is also employed to train the SVM for the broad class classification task, to examine the performance of all the spectral parameters in tandem. The set performs close to the results obtained using parameters M and σ_M. A comparison in classification performance is also attempted using a single layer MLP trained over the set *SF*. The results, as shown in Table 7.2, suggest that the set *SF* gives similar performance with SVM and MLP systems.

The classification within the broad classes is also attempted with a 4–layered end–to–end CNN architecture, with raw cry as input. The training is carried out over a sub-segmental duration with kW=16. The network results in a good classification of 99%, which is better than the parameter set *SF*, given in Table 7.2. The proposed method also uses the source representation obtained using ZFF as input to the CNN architecture. The ZFF input signal also results in a high classification as from the raw signal input.

The task of classification between finer causes of infant cry, such as pain and stranger's anxiety, is also attempted. A set of spectral parameters $\{M, \sigma_M, \Delta M, \sigma_{\Delta M}\}$, obtained over a duration of 30 ms, are used in an MLP architecture, to classify within these classes. A classification score of 78% is obtained. MFCC features gave a high performance in discriminating broader classes of causes of infant cry. A SVM trained on MFCC for the task of classification within the pain and stranger's anxiety classes gives a score of 94%. Table 7.3 shows these results. The classification is further attempted using a 4–layered CNN architecture, with raw cry and ZFF signal as input. The scores obtained over the CNN network show a significant improvement in performance.

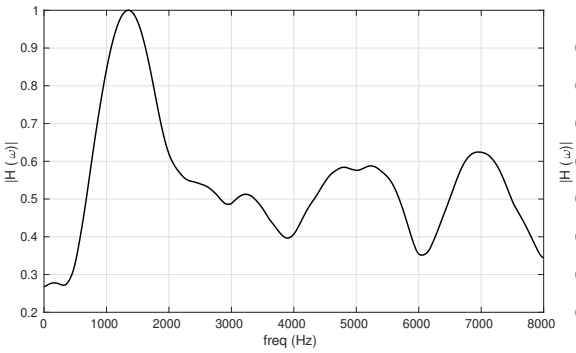


FIGURE 7.3: Gross spectral response obtained from first layer of CNN trained over ZFF signals.

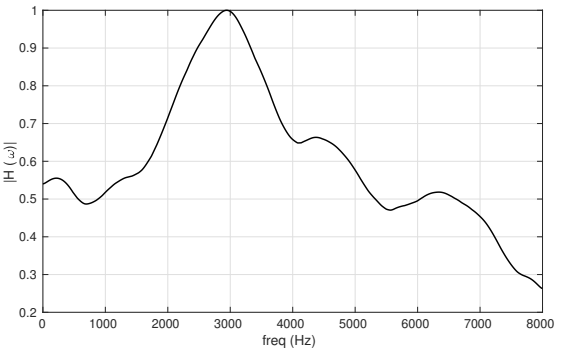


FIGURE 7.4: Gross spectral response obtained from first layer of CNN trained over raw cry signals.

The network with ZFF signal as input gives a higher performance (F–1 = 99%) than raw cry (F–1 = 76%). This reinstates the ability of the ZFF to capture the vocalization characteristics which differentiates between the cry signal behaviour based on the factors causing it. Changing the kW from 16 to 300 does not alter the performance significantly, which shows that the CNNs are able to capture the context related information while learning representations from a smaller segment.

7.4 Analysis

The raw cry and ZFF signals give similar performance for the classification task for broad categories. This shows that the classes are distinct in their source and system behaviour. This, as tabulated in section 5.5.1, Table 5.6, corroborates the observations about the prominent variations in the formant frequency values within a *build-up* stage of a crying bout. Further experiments attempt to classify the finer classes within the severe causes of infant cry, i.e. pain vs. stranger's anxiety.

The first layer in the CNN acts as a band of filters which help in learning an appropriate representation of the input for the classification purpose. An average of the coefficients of all these 128 filters represents a gross spectral response which is being learned by the network as parameters. The parameters are specific to the network architecture and the classification task.

Figure 7.3 and 7.4 show the gross spectral response obtained from the first layer of filters, from 2 layered end–to–end CNN architecture trained over ZFF and raw cry signals respectively. The underlying evaluation is pain vs. stranger's anxiety classification. The response from the network trained over ZFF, shown in Fig. 7.3, shows a high emphasis to frequency components around 1.3 kHz, along with moderate emphasis on components

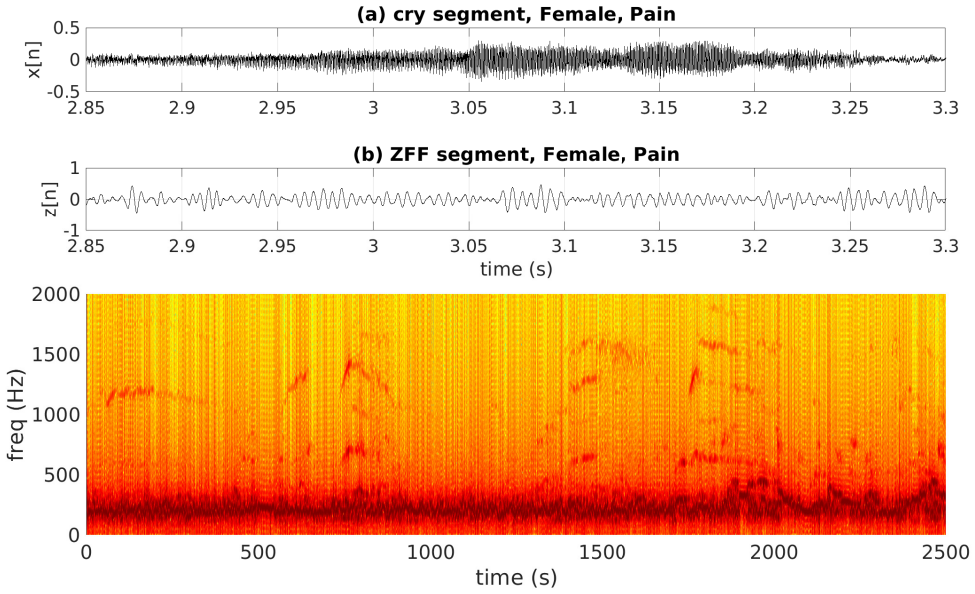


FIGURE 7.5: Speech, ZFF and spectrogram obtained using ZFF for infant cry segment, caused due to pain.

around $3,5$ and 7 kHz. Similarly, the response learned by the CNN trained over cry signals has a prominent peak around 3000 Hz.

An ideal zero frequency signal is expected to contain components only in the vicinity of 0 Hz, and therefore the presence of the higher frequency components in significance raises concerns. This phenomenon can be attributed to the fact that the trend removal in ZFF is not optimal, which results in a presence of high-frequency components in the ZFF spectrum. Figure 7.5 shows this characteristic for a segment of cry signal with the help of the short time spectrogram which reflects presence of significant components between the frequency range 500–1500 Hz. The spread of spectral energy can be seen across the 0–1.5 kHz band in the spectrum. The rapidly changing characteristics and pitch with high value makes it difficult to estimate an optimal duration of the trend removal window in ZFF, which introduces discontinuity in the filtered signal. The gross spectral response is characteristic to the input, network and the task, and varies for a difference in these entities. A similar network trained across adult speech signals for the task of depression detection exhibits a distinct gross spectral response of the filters in first layer [64]. The importance of the spectral feature representation learned in the first layer of CNN network is further verified by boosting the significant components obtained from gross filter response in the input signal. Using the response shown in Fig. 7.3, we design 1–pole and 3–pole resonators, with the location and 3 dB bandwidth estimated from the same response. These resonators are shown in Fig. 7.6 and 7.7. Two resonator systems are designed centered at the first (1300 Hz), and first 3 dominant peaks ($1300, 3100$ and 7100 Hz), to train the network using the input as ZFF signal

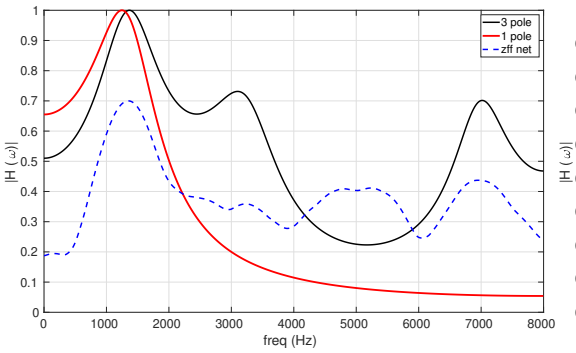


FIGURE 7.6: One pole and 3 pole resonator systems designed to filter the input ZFF signal.

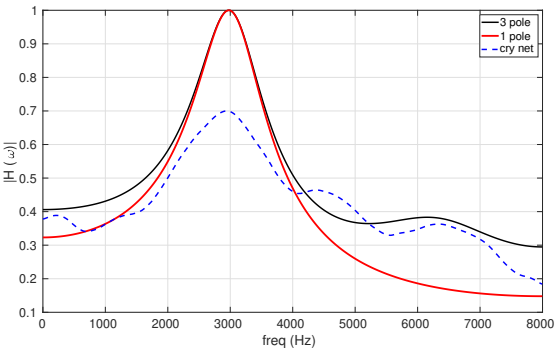


FIGURE 7.7: One pole and 3 pole resonator systems designed to filter the input cry signal.

TABLE 7.4: Performance of filtered signals towards classification of pain and stranger's anxiety cry sounds.

Input	*Filter*	*Precision*	*Recall*	*F*-1
ZFF	1 pole - ZFF net	0.92	0.91	0.91
ZFF	3 poles - ZFF net	0.96	0.96	0.96
Cry	1 pole - Cry net	0.76	0.75	0.74
Cry	3 poles - Cry net	0.76	0.76	0.76

filtered to boost the significant component, as shown in Fig. 7.6. Similar resonators are designed corresponding to the response obtained from the network trained over raw cry signals, and are plotted in Fig. 7.7.

The ZFF signal filtered using the two corresponding resonator systems are used to train the network which give a close yet lower classification performance for the input, as given in Table 7.4. An increase in performance from 92% to 96% is seen when increasing the prominence of first 3 spectral components as compared to one is performed. A similar case is seen for network trained on raw cry signals

7.5 Conclusion

The task of classifying infant cries for different causes is examined by evaluating conventional acoustic features and alternative source representations in this work.

The spectral features set *SF* combined, when evaluated using SVM gives good classification accuracy of 97 %. Features based on MFCCs alone are observed to characterize acoustic properties within cry signals belonging to severe/non-severe categories with good classification accuracies up to 98 %. MFCC based features when evaluated for

classifying between finer categories, pain and strangers anxiety give relatively lower F–1 scores of 94%.

End-to-end learning and classification approach when evaluated over raw cry and ZFF signals are observed to give most optimal performances both at broader and finer level categorization, with best F–1 scores of 99% obtained for each task. The good performance of 99% from both raw cry and ZFF signals indicate the distinctive characterizing capacity of both source and system components within cry signal, towards classification with causes at broader level, whereas the efficacy of source representation, i.e. ZFF is established when classification is performed between the categories at the finer level.

Chapter 8

A Modular Approach towards Learning Baby Sounds Classification using CNNs

The present study focuses on the baby sound sub-challenge of the INTERSPEECH 2019 Computational Paralinguistics ChallengE (ComParE). Towards that, we focus on developing neural networks based raw waveform modeling methods. Speech technology mostly is based upon the fundamentals of speech production mechanism of an adult. In contrast, the proposed approach does not make such assumptions and tries to learn the relevant features from the signal directly.

We investigate two approaches, namely,

1. Monolithic approach: In this approach a single classifier is trained to classify the five sound classes in the data set.
2. Modular approach: In this approach we decompose the classification problem into sub-classification problems. More precisely, we group the five classes into different categories; build sub-classifiers; and finally combine their outputs to predict the sound class.

The modular approach is motivated by the fact that baby sounds tend to exhibit high degree of variability. Furthermore, the amount of the data available is low. So, it is difficult to learn an invariant feature to discriminate all classes through a single end-to-end classifier. Our investigations show that the end-to-end learning/classification approach performs well for a different data-set of infant/baby sounds too, when evaluated

TABLE 8.1: Confusion matrix among all 5 classes with raw speech/ZFF as input.

Target \ *Predicted*	Canonical	Crying	Junk	Laughing	Non-canonical
Canonical	136/86	3/4	93/104	0/0	146/184
Crying	0/3	45/34	21/28	0/0	97/98
Junk	40/34	9/9	1060/1046	0/0	248/268
Laughing	7/7	0/1	12/22	0/0	22/11
Non-canonical	63/83	83/70	345/368	1/0	1186/1157

over the sub-modules of the modular approach. Modular approach leads to better sub-systems towards classifying sounds from all the 5 classes.

8.1 Preliminary Analysis

During training, we follow five-fold cross validation process, where the training data is split into five parts, four parts are used for training and the fifth part is used as cross validation data for early stopping. This results in five classifiers. During testing, we first average the output probabilities per frame across five classifiers and then average across the utterance to make the decision.

8.1.1 Monolithic approach

The monolithic approach uses raw signals modeled by the CNN based architecture for the task of 5 class classification. It can be observed that raw signal lead to better system than ZFF. When compared to the baseline systems, the monolithic approach leads to inferior system on the development set. Performance of the raw waveform or the ZFF signal as input is low compared to the baseline provided with the ComParE challenge results, which is 0.54, obtained using *compare* feature-set along with SVM.

Table 8.1 presents the confusion matrix for the two systems. The counts A/B denote A for raw signal based system and B denotes for ZFF signal based system. It can be observed that in both cases, the classification is high for most frequent classes in the training set i.e. junk class and non-canonical class. In both cases, laugh is never detected.

TABLE 8.2: Unweighted average recall (UAR) for classifiers built over different modules for baby sound classification task.

S. #	*Task details*	*UAR (Cry)*	*UAR (ZFF)*
1	Junk vs. all other classes	*0.77*	*0.77*
2	Canonical vs. non-canonical	*0.69*	0.65
3	Crying vs. Laughing	0.64	*0.77*
4	Crying and Laughing vs. Canonical and Non-canonical	*0.57*	0.52
5	Junk vs. Crying and Laughing	0.72	*0.74*
6	Junk vs. Canonical and Non-canonical	0.77	*0.78*
7	Junk vs. Crying and Laughing vs. Canonical and Non-canonical	*0.56*	–
Average		0.67	0.71

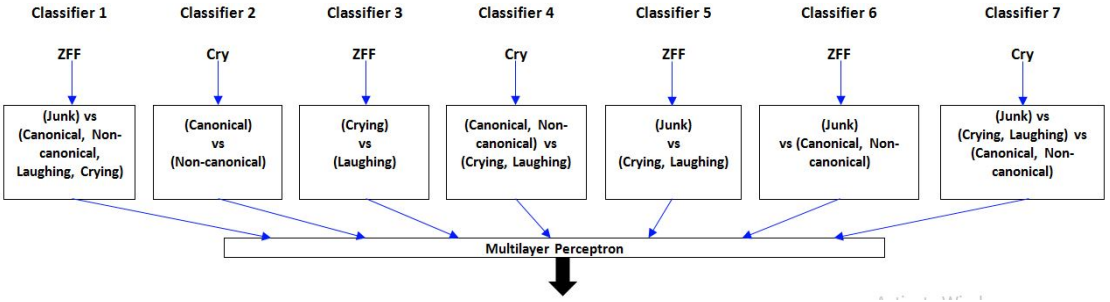


FIGURE 8.1: Proposed modular architecture towards baby sound classification task.

8.2 Results and Analysis

The classes in the training set are grouped towards a divide and conquer approach to the classification problem. The groups are determined by the acoustic content of the classes which can be exploited for an optimal classification. For instance, a comparison of the *canonical vs. non-canonical* shows that a good classification rate can be achieved by using *cry* signals, that helps in modeling both source and system characteristics, present in these categories (see Table 8.2).

8.2.1 Modular approach

The classification task is divided in smaller modules based on the behaviour of signals to improve upon the performance of the monolithic approach. Table 8.2 gives the classification results between different modules, obtained by training 7 different classifiers, depicted schematically in Fig. 8.1. All the classifiers are 4 layered CNN architectures followed by a MLP classifier, and the classification results are interpreted in terms of

TABLE 8.3: Confusion matrix among all 5 classes with inputs as per the modular approach.

Target \ *Predicted*	Canonical	Crying	Junk	Laughing	Non-canonical
Canonical	158	4	80	2	134
Crying	2	45	19	0	97
Junk	57	14	1021	0	265
Laughing	7	2	15	0	17
Non-canonical	99	63	280	2	1234

unweighted average recall *(UAR)* values. A module includes one or more classes grouped as a set to be classified against another set. Decision about the groups is arrived based upon the type of signal within a class. The junk class, for example, comprises mostly of ambient noise, and background babble, which is entirely different from other classes which contain different aspects of the baby sound. The junk class can therefore be classified easily against all other classes together. Hence, the classifier #1 performs this task with both cry and ZFF signal as the input, which suggests that either just speech source or that along with system characteristics are predominant in the group with all 4 classes when compared to junk. Crying and laughing classes comprise of non–speech signals, which are para–linguistic characteristics in the sound. Canonical and non-canonical classes comprise of sounds uttered as speech. The two classes are best classified using the raw speech input signal which gives a classification accuracy of 69% for the classifier #2. Using the ZFF signal gives a lower classification rate which shows that discriminating these classes will also require the knowledge of production system, than just the source characteristics. A good discrimination of 77% is achieved between the crying and laughing classes using the ZFF signal, for the classifier #3. Interestingly the network trained towards classifying crying vs. laughing is trained by freezing the initial 4 layers as that from the prior experiment on pain vs. strangers anxiety classification. This not only helps obtain good classification scores for cry vs. laughing experiments, but also establishes the significance of the spectral components that the CNN learns towards the task of classifying pain and stranger's anxiety cries. Similarly, the performances of other classifiers also can be compared for both the input types from Table 8.2. For classifier #7, the evaluation was done only using cry as input type, considering the presence of categories canonical and non-canonical in the task.

TABLE 8.4: Comparison of performance for different approaches towards baby sounds classification.

Approach	*UAR*
Monolithic (raw)	0.41
Monolithic (ZFF)	0.38
Modular	0.44
Baseline	0.54

8.2.2 Analysis

The confusion matrix observed from the modular approach shown in Table 8.3 is not that different from the one observed from the monolithic approach shown in Table 8.1 for cry input based system. The confusion for classes canonical and non-canonical is observed to be slightly reduced, with increase in their true classification count for the modular approach. The confusion induced for the class laughing, as observed from both the Tables, indicate possible rationality in terms of the predictions made. Most of the predictions are done in favour of classes junk or non-canonical, which the laughing category could possibly have common characteristics with. Whereas, canonical and crying categories have peculiarities with respect to the articulatory variations and signal energy respectively, that laughing sounds severely lack. Category specific generalization could better be achieved for laughing with availability of sufficient training samples.

This brings out the importance of such an approach where different input types are used towards training several task specific end-to-end CNN based classifiers, since no single system is observed to be doing the job well.

The UAR scores from different experiments following monolithic, modular and challenge baseline approaches are compared and the results are shown in Table 8.4. Although very slight, but an improvement of 3 % in the UAR score is observed from the monolithic to modular approach. These results are still below the baseline scores 54 %, which is reported for the system evaluating *compare* features with an SVM.

8.3 Conclusion

In this work, the approach of end-to-end learning and classification using a convolutional neural network is implemented to attempt classification of baby sounds into multiple categories, which are significantly different in terms of their acoustic contents. The data-set used for the multi-class classification task is a collection of different types of baby sounds, obtained as part of INTERSPEECH 2019 baby sounds sub-challenge.

Two approaches involving end-to-end CNNs, *monolithic* and *modular* approaches are examined towards multi-class classification. Monolithic approach involves classification of baby sounds into 5 pre-defined classes using a single CNN. A modular approach is designed to divide the problem into smaller sub-problems of classifying baby sounds into categories defined from the given category-set, based upon the signal content the categories represent. As against the usage of a single input type of either raw cry or zero frequency filtered signal, for monolithic approach, the input in modular approach is decided based upon the signal content, the categories involved in an experiment represent.

Junk category is observed to be best represented by ZFF signals. The canonical and non-canonical categories, representing sounds that have subtle signatures of evolving speech characteristics are optimally characterized using raw cry signals. Raw cry signal, which consists information of both excitation source and system characteristics and therefore effectively helps with the classification into these categories. Laughing and crying category sounds are optimally classified using the ZFF representation of the cry signal. The performance of this sub-system is observed to increase significantly, when the knowledge learned from the other similar tasks like for pain vs. stranger's anxiety classification, for infant cries, is transferred for its training purpose. This effectively establishes the significance of the spectral cues learned by a CNN trained to classify cries between different causes.

The confusion reported from the performance of the experiments in this work is observed to be least for the categories that have significantly high count of training data i.e., junk and non-canonical, followed by canonical and crying categories, those have relatively higher classification confusion. Laughing category sounds could not be classified when evaluated as part of the overall multi-class classification setup, by any approach examined in the work. A prominent reason being significantly scarce data for this category. An approach like end-to-end classification, that critically depends upon the availability of sufficient per-category training data, is not able to achieve results comparable to that of the base-line that effectively evaluates a comprehensive set of 6373 static features, with SVM. The performance obtained from the modular approach is observed to be better than that for monolithic approach by 3%, but is still significantly below the sub-challenge baseline score of 54%. Nevertheless, additional pre-processing of the baby-sounds along with increasing the training data for laughing category could possibly help make the systems performance comparable to that of the baseline system.

Chapter 9

Summary and Conclusion

In this manuscript, acoustic characteristics of infant cry sound production are explored for the characterization of the cry-causes and different cry sounds sub-types, along with evaluation of automated crying/vocalization type recognition. Such insights could be useful in further analyzing acoustics in para–linguistic and non-verbal sounds. Significant differences are observed between the acoustic features for severe and non-severe cry-cause categories along with the ones at the finer level of categorical granularity, using f_0 contour, sub-band spectral energy ratios and formant frequencies $F_1 - F_5$. Cry sub-type characterization proposed here may be useful to extract information such as age, pathological conditions and cry-cause severity. An attempt is made towards deriving insights into the cry-cause categories and cry sound sub-types, using a constrained data-set.

Evaluation of automated infant vocalization cause/category recognition systems has been done by first assessing spectral features based on f_0, sub-band spectral energy and MFCC and their derivatives. The classification performance for discriminating severe and non-severe cause cry sounds, is observed to be good for MFCC and their derivatives when evaluated using SVM and multilayer perceptron, with 0.98 and 0.96 F-1 scores respectively. Their performance drops down to 0.94 and 0.78 respectively when the classification is examined at the finer level of categorical granularity, i.e., for pain and strangers anxiety. A relatively recent approach of end-to-end learning and classification using CNN is evaluated for the same tasks. As inputs, raw cry and zero frequency filtered signals are used. An overall improvement in the classification performance is observed. Good F-1 scores of 0.99 each, for the CNN evaluated over the raw cry and ZFF signals for classifying severe and non-severe cries are obtained. Whereas, this approach is observed to yield comparatively better for pain vs. stranger's anxiety classification with 0.99 F-1 score as compared to that from spectral features, with conventional machine learning

or MLP based techniques. As against the conventional approach of first deriving hand-crafted features and then performing classification, the approach of learning suitable acoustic descriptors and classifying, in tandem, is found to help model the characteristics that represent a highly varying entity of a signal as that of an infant's vocalization.

The approach of end-to-end learning using a CNN is also implemented to assess the task of classification of baby sounds into categories canonical, crying, junk, laughing and non-canonical. A *monolithic* approach that implements a CNN in a manner same as implemented for the cry-cause classification tasks is explored. Another approach called *modular* approach that requires division of the 5-class classification task into sub-tasks with lesser categories, followed by a common training of an MLP towards the 5-class classification is also comparatively examined. The *modular* design of these classification tasks and their corresponding inputs are based on the type of the acoustic contents different categories represent. For instance, canonical and non-canonical represent categories having speech-like traits in the sounds. Whereas, crying and laughing do not involve significant articulatory variations. Instead, they are mostly characterized by highly varying, primary and alternative excitation patterns, which are effectively captured by the zero frequency filtering. Modular approach is observed to serve the purpose better with 0.44 unweighted average recall (UAR) as compared to 0.41 in case of monolithic approach. Although, significantly lesser than the baseline UAR of 0.54 reported in the challenge paper, a slight improvement in the classification scores using modular approach over the monolithic approach suggest its effectiveness in cases where one approach/input type-based solution does not suffice for an optimal performance.

The implications for the utility of such technology are of immense value. Early detection of the medical conditions of infants in cases where its not that simple to know the cause can be crucial for a better health-care. Thus making communication between an infant and his/her surrounding more meaningful. The infant cry analysis needs to happen for the data which is sufficient and as real-time as possible, specifically for the categories like ailment or hunger/thirst, where the data-set is hard to procure. A solution could be crowd-sourcing of real-time cry data collection by having parents/guardians record cries of their babies with the help of a mobile app and comprehensively annotate it.

The analytical observations made in this work still need better validation checks, corresponding to more reliable ground-truth based infant cry data-set. The aspects of cry acoustics discussed in this work along with better and larger data-set, especially that involving multiple recordings by same infants over multiple sessions and for different causes could help provide better insights for studies concerned with excitation source and production system characteristics of infant cry sound, while accounting subtle nuances crucial for the tasks related to cry sound recognition for cause based analysis.

Chapter 10

List of Publications

The following is the list of publications that are written based on the research work done.

1. Shivam Sharma, P. Viswanth and Vinay Kumar Mittal, "Infant Crying Cause Recognition using Conventional and Deep Learning based Approaches", in *Proceedings of the 15th International Conference on Natural Language Processing (ACL)*, Dec. 15, 2018.
   Details: This paper presents an attempt towards performing cry sound classification, into severe and non-severe categories. The work presented explores through the possibilities offered by algorithms like SVM, multilayer perceptron and convolutional neural networks towards this task. The classification performance obtained is encouraging and comparable with other similar State-of-the-Art implementations reported as part of the literature.

2. S. Sharma and V. K. Mittal, "Infant cry analysis of cry signal segments towards identifying the cry-cause factors," in *TENCON 2017 IEEE Region 10 Conference*, Penang, 2017, pp. 3105-3110. doi: 10.1109/TENCON.2017.8228395
   Details: As part of this work, distinctive analysis of the acoustic characteristics of infant cry sounds is performed, based upon basic statistical parameters like mean, standard-deviation and normalized standard-deviation of f_0 contour.

3. Shivam Sharma, Pruthvi Raj Myakala, Rajasree Nalumachu, Suryakanth V. Gangashetty and V. K. Mittal, "A Study on Acoustic Features of Infant Cry Signal for Different Causes of Crying", in 3^{rd} *Int. Workshop on Affective Social Multimedia Computing (ASMMC) 2017*, Co-located with INTERSPEECH, Stockholm, Sweden, Aug. 25th, 2017.
   Details: This work reports the details of the primary acoustic analysis performed

on IIIT-S ICSD2 data-set, conducted using f_0 contour, signal energy and formant frequencies.

4. S. Sharma, P. R. Myakala, R. Nalumachu, S. V. Gangashetty and V. K. Mittal, "Acoustic analysis of infant cry signal towards automatic detection of the cause of crying," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, San Antonio, TX, 2017, pp. 117-122. doi: 10.1109/ACIIW.2017.8272600
   Details: This paper extends on the observations made in the work on acoustic analysis, but with other infant subjects. Also the formant analysis is done using a different approach.

5. S. Sharma and V. K. Mittal, "A qualitative assessment of different sound types of an infant cry," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, Mathura, 2017, pp. 532-537. doi: 10.1109/UPCON.2017.8251106
   Details: The phenomenon and observations related to different dysphonatory sounds present within an infant cry, were first acknowledged as part of this work. It also discusses the correlation between different dysphonation types and various factors like infant's age or pathological conditions.

6. Shivam Sharma, Shubham Asthana, and V. K. Mittal. "A database of infant crysounds to study the likely cause of cry", in *Proc. of the 12th International Conference on Natural Language Processing*, Trivandrum, India, NLP Association of India, December 2015, pp. 112117.
   Details: IIIT-S ICSD was first introduced through this publication. It also builds upon preliminary acoustic analysis of cry signals for pain and discomfort categories.

Others

1. P. R. Myakala, R. Nalumachu, *S. Sharma* and V. K. Mittal, "A low cost intelligent smart system for real time infant monitoring and cry detection," in *TENCON 2017 IEEE Region 10 Conference,* Penang, 2017, pp. 2795-2800. doi: 10.1109/TENCON.2017.8228337
   Details: The work in this paper extends on the prototype, built as part of an earlier attempt to facilitate real-time infant cry monitoring. This work considers the analysis and detection of pain and discomfort cry sounds.

2. P. R. Myakala, R. Nalumachu, *S. Sharma* and V. K. Mittal, "An intelligent system for infant cry detection and information in real time," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops*

and Demos (ACIIW), San Antonio, TX, 2017, pp. 141-146.
doi: 10.1109/ACIIW.2017.8272604

Details: The paper reports on our first attempt towards prototyping a hard-ware device that can help monitor the crying activity and facilitate notification based reaction mechanism. Primary cause-category considered in this work is pain.

Bibliography

[1] Bjorn Schuller, Anton Batliner, Christian Bergler, Florian B. Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, Alejandrina Cristi, Amanda Seidl, Anne Warlaumont, Lisa Yankowitz, Elmar Nth, Shahin Amiriparian, Simone Hantke, and Maximilian Schmitt. The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *INTERSPEECH*. ISCA, 2019. URL http://www.compare.openaudio.eu/wp-content/uploads/2019/03/INTERSPEECH_2019_ComParE.pdf.

[2] John H Esling. The articulatory function of the larynx and the origins of speech. In *Annual Meeting of the Berkeley Linguistics Society*, volume 38, pages 121–149, 2012.

[3] Hemant A Patil. cry baby: Using spectrographic analysis to assess neonatal health status from an infants cry. In *Advances in speech recognition*, pages 323–348. Springer, 2010.

[4] John H Esling. Pharyngeal consonants and the aryepiglottic sphincter. *Journal of the International Phonetic Association*, 26(2):65–88, 1996.

[5] John H. Esling and Jerold A. Edmondson. *The Laryngeal Sphincter as an Articulator: Tenseness, Tongue Root and Phonation in Yi and Bai.* A. Braun and H. R. Masthoff, eds., Phonetics and its Applications, Stuttgart: Franz Steiner Verlag, 2001. Festschrift for JensPeter Kster on the Occasion of his 60th Birthday.

[6] John H. Esling, Katherine E. Fraser, and Jimmy G. Harris. Glottal stop, glottalized resonants, and pharyngeals: A reinterpretation with evidence from a laryngoscopic study of nuuchahnulth (nootka). *Journal of Phonetics - J PHONETICS*, 33:383–410, 10 2005. doi: 10.1016/j.wocn.2005.01.003.

[7] Jerold A. Edmondson, Ccile M. Padayodi, Zeki Majeed Hassan, and John H. Esling. The laryngeal articulator: Source and resonator.

[8] John H Esling. There are no back vowels: The larygeal articulator model. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 50(1-4):13–44, 2005.

[9] D Crystal. In: Sebeok, t. (ed.), paralinguistics. In *Current Trends in Linguistics*, volume 12, page 265295, The Hague, 1974. Mouton.

[10] Linda L. LaGasse, A. Rebecca Neal, and Barry M. Lester. Assessment of infant cry: Acoustic cry analysis and parental perception. *Mental Retardation and Developmental Disabilities Research Reviews*, 11(1):83–93, 2005.

[11] C Manfredi, L Bocchi, S Orlandi, L Spaccaterra, and GP Donzelli. High-resolution cry analysis in preterm newborn infants. *Medical engineering & physics*, 31(5): 528–532, 2009.

[12] Amy Neustein, editor. *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*. Springer, New York, 2010. ISBN 978-1-4419-5950-8. doi: 10.1007/978-1-4419-5951-5.

[13] M. Petroni, M.E. Malowany, C.C. Johnston, and B.J. Stevens. A new, robust vocal fundamental frequency (f0) determination method for the analysis of infant cries. In *Computer-Based Medical Systems, 1994., Proceedings 1994 IEEE Seventh Symposium on*, pages 223–228, Jun 1994. doi: 10.1109/CBMS.1994.316016.

[14] Rami Cohen and Yizhar Lavner. Infant cry analysis and detection. In *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pages 1–5. IEEE, 2012.

[15] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman. Baby cry detection in domestic environment using deep learning. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, pages 1–5, Nov 2016. doi: 10.1109/ICSEE.2016.7806117.

[16] Bjorn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 312–315. ISCA, 2009. URL http://dblp.uni-trier.de/db/conf/interspeech/interspeech2009.html#SchullerSB09.

[17] Bjorn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus R. Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In Frdric Bimbot, Christophe Cerisara, Ccile Fougeron, Guillaume Gravier, Lori Lamel,

Franois Pellegrino, and Pascal Perrier, editors, *INTERSPEECH*, pages 148–152. ISCA, 2013. URL http://dblp.uni-trier.de/db/conf/interspeech/interspeech2013.html#SchullerSBVSRCWEMMSPVK13.

[18] Vinay Kumar Mittal and B Yegnanarayana. Effect of glottal dynamics in the production of shouted speech. *The Journal of the Acoustical Society of America*, 133 (5):3050–3061, 2013.

[19] Vinay Kumar Mittal and B Yegnanarayana. Analysis of production characteristics of laughter. *Computer Speech & Language*, 30(1):99–115, 2015.

[20] Vinay Kumar Mittal and B Yegnanarayana. Study of characteristics of aperiodicity in Noh voices. *The Journal of the Acoustical Society of America*, 137(6):3411–3421, 2015.

[21] Ada Fort and Claudia Manfredi. Acoustic analysis of newborn infant cry signals. *Medical Engineering & Physics*, 20(6):432 – 442, 1998.

[22] A. Fort, A. Ismaelli, C. Manfredi, and P. Bruscaglioni. Parametric and non-parametric estimation of speech formants: application to infant cry. *Medical Engineering & Physics*, 18(8):677 – 691, 1996.

[23] Vinay Kumar Mittal. Discriminating the infant cry sounds due to pain vs. discomfort towards assisted clinical diagnosis. In *Proc. SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies*, pages 37–42, 2016. doi: 10.21437/SLPAT.2016-7. URL http://dx.doi.org/10.21437/SLPAT.2016-7.

[24] V. K. Mittal. Discriminating features of infant cry acoustic signal for automated detection of cause of crying. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5, Oct 2016. doi: 10.1109/ISCSLP.2016.7918391.

[25] Hesam Farsaie Alaie and Chakib Tadj. Cry-based classification of healthy and sick infants using adapted boosting mixture learning method for gaussian mixture models. *Modelling and Simulation in Engineering*, 2012:55, 2012.

[26] Yasmina Kheddache and Chakib Tadj. Identification of diseases in newborns using advanced acoustic features of cry signals. *Biomedical Signal Processing and Control*, 50:35–44, 2019.

[27] Lina Abou-Abbas, Hesam Fersaie Alaie, and Chakib Tadj. Automatic detection of the expiratory and inspiratory phases in newborn cry signals. *Biomedical Signal Processing and Control*, 19:35 – 43, 2015.

[28] Susan M. Grau, Michael P. Robb, and Anthony T. Cacace. Acoustic correlates of inspiratory phonation during infant cry. *Journal of Speech, Language, and Hearing Research*, 38(2):373–381, 1995.

[29] Robert F. Orlikoff, R. J. Baken, and Dennis H. Kraus. Acoustic and physiologic characteristics of inspiratory phonation. *The Journal of the Acoustical Society of America*, 102(3):1838–1845, 1997.

[30] Chuan-Yu Chang, Chuan-Wang Chang, S. Kathiravan, Chen Lin, and Szu-Ta Chen. Dag-svm based infant cry classification system using sequential forward floating feature selection. *Multidimensional Systems and Signal Processing*, 28(3):961–976, 2017.

[31] Jen Hirschberg. Dysphonia in infants. *International Journal of Pediatric Otorhinolaryngology*, 49:S293 – S296, 1999. ISSN 0165-5876. doi: https://doi.org/10.1016/S0165-5876(99)00179-2. URL http://www.sciencedirect.com/science/article/pii/S0165587699001792.

[32] Yasmina Kheddache and Chakib Tadj. Acoustic measures of the cry characteristics of healthy newborns and newborns with pathologies. *Journal of Biomedical Science and Engineering*, 6(08):796, 2013.

[33] Maria A. Proytcheva. Issues in neonatal cellular analysis. *American Journal of Clinical Pathology*, 131(4):560–573, 2009. doi: 10.1309/AJCPTHBJ4I4YGZQC. URL http://dx.doi.org/10.1309/AJCPTHBJ4I4YGZQC.

[34] Yasmina Kheddache and Chakib Tadj. Resonance frequencies behavior in pathologic cries of newborns. *Journal of Voice*, 29(1):1 – 12, 2015. ISSN 0892-1997. doi: https://doi.org/10.1016/j.jvoice.2014.04.007. URL http://www.sciencedirect.com/science/article/pii/S0892199714000757.

[35] Marco Cecchini, Carlo Lai, and Viviana Langher. Dysphonic newborn cries allow prediction of their perceived meaning. *Infant Behavior and Development*, 33(3):314 – 320, 2010. ISSN 0163-6383. doi: https://doi.org/10.1016/j.infbeh.2010.03.006. URL http://www.sciencedirect.com/science/article/pii/S0163638310000469.

[36] Takeo Fujiwara, Ronald G. Barr, Rollin Brant, and Marilyn Barr. Infant distress at five weeks of age and caregiver frustration. *The Journal of Pediatrics*, 159(3):425 – 430.e2, 2011. ISSN 0022-3476. doi: https://doi.org/10.1016/j.jpeds.2011.02.010. URL http://www.sciencedirect.com/science/article/pii/S0022347611001582.

[37] Philip Sanford Zeskind and Elisabeth A. Shingler. Child abusers' perceptual responses to newborn infant cries varying in pitch. *Infant Behavior and Development*, 14(3):335 – 347, 1991. ISSN 0163-6383. doi: https://doi.org/10.1016/0163-6383(91)90026-O. URL http://www.sciencedirect.com/science/article/pii/0163638391900260.

[38] A. Chittora and H. A. Patil. Classification of normal and pathological infant cries using bispectrum features. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 639–643, Aug 2015. doi: 10.1109/EUSIPCO.2015.7362461.

[39] N. S. A. Wahid, P. Saad, and M. Hariharan. Automatic infant cry classification using radial basis function network. In *Journal of Advanced Research in Applied Sciences and Engineering Technology*, volume 4 of *1*, pages 12–28, 2016. ISSN (online): 2462-1943.

[40] S Chandralingam, T Anjaneyulu, and K Satyanarayana. Estimation of fundamental and formant frequencies of infants cries; a study of infants with congenital heart disorder.

[41] Shubham Asthana, Naman Varma, and Vinay Mittal. Preliminary analysis of causes of infant cry. pages 000468–000473, 12 2014.

[42] Shivam Sharma, Shubham Asthana, and V. K. Mittal. A database of infant cry sounds to study the likely cause of cry. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 112–117, Trivandrum, India, December 2015. NLP Association of India. URL http://www.aclweb.org/anthology/W/W15/W15-5917.

[43] Shivam Sharma, Pruthvi R. Mykala, Rajasree Nalumachu, Suryakanth V. Gangashetty, and Vinay Kumar Mittal. A study on acoustic features of infant cry signal for different causes of crying. *3rd international workshop on Affective Social Multimedia Computing (ASMMC), INTERSPEECH 2017*, 2017. Stockholm.

[44] Anshu Chittora and Hamant A Patil. Data collection and corpus design for analysis of nonnal and pathological infant cry. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–6. IEEE, 2013.

[45] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time signal processing*, volume 2. Prentice-hall Englewood Cliffs, 1989. pp. 71.

[46] Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*, volume 14, page 123.

[47] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4): 1917–1930, 2002.

[48] James Lyons. Mel frequency cepstral coefficient (mfcc) tutorial, Jan 2012. URL http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

[49] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63 (4):561–580, April 1975. ISSN 0018-9219. doi: 10.1109/proc.1975.9792. URL http://dx.doi.org/10.1109/proc.1975.9792.

[50] Chang, Chang Chuan-Yu, Chuan-Wang, Kathiravan, Lin S, Chen, Chen, and Szu-Ta. Dag-svm based infant cry classification system using sequential forward floating feature selection. *Multidimensional Systems and Signal Processing*, 28(3):961–976, 2017.

[51] K Sri Rama Murty and B Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613, 2008.

[52] Simon Haykin. *An Introduction to Analog and Digital Communications*. John Wiley & Sons, Inc., New York, NY, USA, 1989. ISBN 0-471-85978-8. pp. 652.

[53] Hervé Bourlard and Stéphane Dupont. Subband-based speech recognition. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1251–1254. IEEE, 1997.

[54] B. Yegnanarayana and K. Sri Rama Murty. Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):614–624, M 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2012194.

[55] John G Proakis. *Digital signal processing: principles algorithms and applications*. Pearson Education India, fourth edition, 2001.

[56] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. ISSN 1573-0565. doi: 10.1023/A:1022627411411. URL https://doi.org/10.1023/A:1022627411411.

[57] djmw. Feedforward neural networks 1. what is a feedforward neural network?, April 2004. URL http://www.fon.hum.uva.nl/praat/manual/Feedforward_neural_networks_1__What_is_a_feedforward_ne.html.

[58] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, Aug 2018. ISSN 1869-4101. doi: 10.1007/s13244-018-0639-9. URL https://doi.org/10.1007/s13244-018-0639-9.

[59] Google Cloud Platform. Best practices, Feb 2018. URL https://cloud.google.com/speech/docs/best-practices.

[60] Mike Brookes. Voicebox: Speech processing toolbox for matlab, oct 2015. URL http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[61] Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition. *Speech Communication*, 2019.

[62] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai Doss. On learning to identify genders from raw speech signal using cnns. Technical report, 2018.

[63] S Pavankumar Dubagunta, Selen Hande Kabil, and Mathew Magimai Doss. Improving children speech recognition through feature learning from raw speech signal. Technical report, 2019.

[64] S Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai Doss. Learning voice source related information for depression detection. Technical report, 2019.

[65] P Gangamohan, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, and B Yegnanarayana. Excitation source features for discrimination of anger and happy emotions. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[66] Sudarsana Reddy Kadiri, P Gangamohan, Suryakanth V Gangashetty, and Bayya Yegnanarayana. Analysis of excitation source features of speech for emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.