Crying For a Reason

A Signal Processing Based Approach for Infant Cry Analysis and Classification

by

Anshu Chittora

201021012

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in

INFORMATION AND COMMUNICATION TECHNOLOGY

to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND
COMMUNICATION TECHNOLOGY
Gandhinagar, India

January, 2017

Declaration

I hereby declare that the thesis comprises of my original work towards the degree of Doctor of Philosophy in Information and Communication Technology (DA-IICT) at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

i. due acknowledgment has been made in the text to all the reference material used.

Anshu Chittora

(ID: 201021012)

Certificate

This is to certify that the thesis work entitled "Crying for a Reason: A Signal Processing Based Approach for Infant Cry Analysis and Classification" has been carried out by Anshu Chittora for the degree of Doctor of Philosophy in information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) under my supervision.

Dr. Hemant A. Patil

Thesis Supervisor

A tribute to my Father

Late Sh. B.L. Chittora

Acknowledgements

This thesis would not have been possible without the help of many people who have been there to support me directly or indirectly. I thank all of them for being beside me.

First of all, I would like to thank my guide Prof. (Dr.) Hemant A. Patil, for giving me guidance and his precious time. He has always been supportive to me during my Ph. D. work. I thank him for providing me all resources necessary for research work and well equipped Speech Research Laboratory. I thank DA-IICT for supporting me and providing me necessary resources.

I would like to thank Department of Electronics and Information Technology (DeitY), New Delhi for sponsoring the consortium projects "Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages" and "Development of Text-To-Speech (TTS) synthesis system in Indian Languages: High Quality Text-to-Speech Synthesis and Small Footprint TTS Integrated with Disability Aids". I would also like to acknowledge Department of Science and Technology (DST), New Delhi for sponsoring the project "Development of Infant Cry Analyzer using Source and System Features". These projects have helped me in getting financial support for several publications.

I would like to thank members of my Research Progress Seminar (RPS) committee and thesis examination committee members, Prof. (Dr.) R. Nagaraj, Prof. (Dr.) M.V. Joshi, Prof. (Dr.) Asim Banerjee, and Prof. (Dr.) Sanjeev Gupta for their precious feedback on my thesis work. I avail this opportunity to express my sincere thanks to Prof. (Dr.) Aditya Tatu, Ph. D. Coordinator, for ensuring smooth evaluation of the reaserch work and providing necessary directions towards completion of the thesis.

I would like to thank Prof. (Dr.) R. Nagaraj, Director, DA-IICT for providing travel grant for presenting my research papers during an international conference at Kuala Lumpur, Malaysia, where I received the best paper award.

I extend my thanks to M. Tech. Coordinator, Prof. (Dr.) V. Sunitha, for providing me TA duty of my choice and staff of resource center, for providing me necessary resources as and when required.

I thank all the members of speech research lab who have helped me in this thesis. My special thanks to Tanvina, Maulik, Hardik, Nirmesh, Kewal, and all those who have helped me directly or indirectly during this research work.

I would like to thank all those parents and hospital authorities of B. J. Medical College, Ahmedabad, India, for supporting me during data collection without which this research would have not been possible. I thank all the professors and medical practitioners of B. J. Medical College for supporting me in data collection work. I would like to thank Prof. Neelam Raval and Prof. Menariya, Head Pediatric Ward, Civil Hospital, for giving me permission for data collection. I express my heartiest thank to Ms. Neeharika Buddha for providing such a rich database which helped a lot to me in my research work. In particular, Ms. Neeharika collected data from three hospitals, namely, King George Hospital, Visakhapatnam, Prabha Nursing Home, Visakhapatnam and Child Clinic, Visakhapatnam. I would like to thank Dr. Carlos A. Reyes-Garcia, Dr. Emilio Arch-Tirado and his INR-Mexico group, and Dr. Edgar M. Garcia-Tamayo for their efforts towards the collection of the infant cry database and providing their database to DA-IICT.

I express my heartiest thanks to my family members especially, my husband, Mr. Jasmit Shah, for supporting me in doing this Ph. D. work and encouraging me whenever I get depressed. I express my gratitude towards my mother-in-law and father-in-law, for supporting me for further studies and taking care of my family responsibilities. I pray to God that every lady should get such supporting in-laws. This thesis would not have been possible without the blessings of my mother, sisters, brother and all the relatives.

I cannot forget to acknowledge the contributions of my daughter, Ms. Tanishi Shah. I always feel sorry for her because I could not give her time which she deserves. My son, Mr. Vardan, is also part of this story and I appreciate that he also supported me refreshing me with his actions and activities. Because of my goals, they could not enjoy their childhood as they could do.

I cannot forget the support given by Ms. Ganga Maheshwari who has taken care of my kids during my Ph. D. work. She has helped a lot in providing safety and bringing up of my kids.

At last, I would like to thank my best companion during the research work, my laptop (Dell IN 4070) which has always supported me and helped me in getting my research work done in time.

Table of Contents

Declaration i

Certificate i

Acknowledgements iii

Table of Contents vi

Abstract.. xiv

List of Principal Symbols xvi

List of Acronyms xvii

List of Tables xviii

List of Figures xxiv

Chapter 1. 1

Introduction 1

1.1 What is an Infant's Cry? 1

1.2 Motivation 2

1.3 Social Relevance of this Research Work..4

1.4 Applications 5

1.5 Research Issues in Infant Cry Analysis 7

1.6 Contributions in the Thesis 8

1.7 System Architecture...10

1.8 Organization of the Thesis 12

1.9 Chapter Summary 14

Chapter 2. 15

Evolution of Infant Cry Analysis and Methods..15

2.1 Introduction 15

2.2 Why Infants Cry?....18
2.3 The Physiology of Cry and Speech Signal....19
2.4 Physioacoustic Model of Cry Production....20
2.5 Organization of Central Nervous System (CNS) in Infants....23
2.6 Factors Affecting Crying Behaviour....24
2.7 Infant Cry Analysis Techniques....25
  2.7.1 Auditory Analysis....25
  2.7.2 Time-Domain Analysis....26
  2.7.3 Frequency-Domain Analysis....27
  2.7.4 Spectrographic Analysis....27
2.8 Recent Trends in Infant Cry Analysis....29
2.9 Literature Search on Sudden Infant Death Syndrome (SIDS)....37
2.10 Chapter Summary....39
Chapter 3.....40
Data Collection and Corpus Design....40
3.1 Introduction....40
3.2 Need for Infant Cry Database or Corpus....40
3.3 Ethical Issues in Infant Cry Data Collection and Protection of Human Rights....42
3.4 Metadata Preparation....44
3.5 Ideal Characteristics of Infant Cry Corpus....45
3.6 Efforts and Experiences during Data Collection....46
3.7 Database Collected and Used in the Research....47
3.8 Factors Influencing Infant Cry Characteristics....51
  3.8.1 Variability of Acoustic Features with Age (newborn to *1* year)...51

3.8.2 Variability of Acoustic Features with Weight of the Infants 55

3.8.3 Cry Type Differences 56

3.8.4 Anatomical Differences in Airways 57

3.8.5 Variability of Acoustic Features with Gestation Age (GA) in Premature Infants 58

3.8.6 Variability of Acoustic Features with Pathologies 58

3.8.7 Birth Cry or First Cry Analysis 61

3.8.8 Stimulation for Cry Production 61

3.8.9 Room Acoustics 62

3.8.10 Time and Duration of Infant Cries 62

3.8.11 Recording Device 62

3.9 Chapter Summary 64

Chapter 4. 66

Signal Processing Challenges in Infant Cry Analysis 66

4.1 Introduction 66

4.2 Short-Time Fourier Transform (STFT) Analysis of Speech Signals 67

4.3 Linear Prediction (LP) of the Speech 71

4.4 Analysis of Results with respect to Spectral Matching 74

4.4.1 Effect of Order of Linear Predictor 75

4.5 Cepstral Analysis on Infant Cry Signal 84

4.6 TEO Analysis of the Infant Cry Signal 90

4.7 Chapter Summary 95

Chapter 5. 97

Analysis of Infant Cries 97

5.1 Introduction 97

5.2 Cryunit Segmentation....98

5.3 Spectrographic Analysis of Infant Cries....99

5.3.1 Introduction of Short-Time Fourier Transform (STFT) Analysis.99

5.3.2 Selection of Window Length....104

5.3.3 Selection of Window Function....106

5.3.4 Infant Cry Modes from Narrowband Spectrogram....109

5.3.5 Spectrographic Analysis and Observations....111

5.3.6 Summary of Observations from Spectrographic Analysis....122

5.4 Classification of Adult Speech and Infant Cry Signals....125

5.4.1 Experimental Results....127

5.4.2 Conclusions....130

5.5 Analysis of Normal and Pathological Infant Cries....130

5.5.1 Preprocessing and Feature Extraction....130

5.5.2 Autocorrelation Method....131

5.5.3 Experimental Results....132

5.6 Newborn's Cry Analysis....136

5.6.1 Estimation of Pitch (F_0) Contour....137

5.6.2 Feature Extraction....140

5.6.3 Experimental Results....141

5.6.4 Analysis using Fundamental Frequency (F_0)-Based Features:...142

5.6.5 Analysis using Normalized Energy-based features....144

5.6.6 Analysis Using Unvoicing Ratio of the Cry....148

5.6.7 Analysis of Results....148

5.7 Significance of Unvoiced Segments and Fundamental Frequency (F_0) in Infant Cry Analysis....150

5.7.1 Teager Energy Operator (TEO)-Based F_0 Estimation Algorithm151

5.7.2 Experimental Setup and Results 155

5.7.3 Analysis of Results Obtained 158

5.8 Sudden Infant Death Syndrome (SIDS) 159

5.8.1 Cry Characteristics of SIDS Victims 161

5.8.2 Respiratory Instability and SIDS Siblings 162

5.9 Analysis of Infant Cries of Respiratory Distress and Infants at Risk.164

5.10 Chapter Summary 166

Chapter 6. 168

Classification of Normal and Pathological Infant Cries 168

6.1 Introduction 168

6.2 Higher-Order Spectral Analysis (HOSA) 169

6.3 Bispectrum Estimation 172

6.3.1 Indirect Method...172

6.3.2 Direct Method..173

6.4 Feature Extraction from Bispectrum 174

6.4.1 Method A- Using Triangular Symmetry of Bispectrum..............174

6.4.2 Method B- Peaks, Peak Locations and Entropy............................175

6.4.3 Method C- Diagonal Slice of Bispectrum 176

6.4.4 Higher-Order Singular Value Decomposition (HOSVD) 176

6.5 Performance Measures 178

6.6 Classification of Normal and Pathological Infant Cries 180

6.6.1 Database 180

6.6.2 Experiment -1...180

6.6.3 Experiment-2 183

6.6.4 Experiment -3 185

6.7 Classification of Pathological Infant Cries 200

6.8 Classification of Normal, Asthma and HIE Infant Cries 202

6.8.1 Database 202

6.8.2 Feature Extraction 203

6.8.3 Experimental Results 203

6.9 Modulation Spectrogram-Based Features 206

6.9.1 Modulation Spectrogram 206

6.9.2 Experimental Setup 208

6.9.3 Experiment 1 208

6.9.4 Experiment 2 209

6.9.5 Experiment 3 209

6.9.6 Experimental Results 209

6.9.7 Summary of Results 212

6.10 Glottal Inverse Filtering (GIF)-Based Features 213

6.10.1 Glottal Waveform Estimation 214

6.10.2 Feature Extraction 215

6.10.3 Experimental Setup and Results 216

6.10.4 Summary of Results 219

6.11 Auditory Spectrogram-Based Features 220

6.11.1 Auditory Spectrogram *vs.* Short-Time Fourier Transform Spectrogram (STFT) 223

6.11.2 Feature Extraction 224

6.11.3 Experimental Setup and Results 225

6.11.4 Summary of Results 229

6.12 Group Delay-Based Features 230

6.12.1 Modified Group Delay Function 230

6.12.2 Feature Extraction 231

6.12.3 Database Used 232

6.12.4 Experimental Results 232

6.12.5 Summary of Results 237

6.13 Classification of Normal *vs*. Deaf Infant's Cries 238

6.13.1 Experimental Setup and Results 238

6.14 Chapter Summary 240

Chapter 7. 242

Conclusions and Future Work 242

7.1 Conclusions 242

7.2 Scope of Future Research Directions 245

7.2.1 Signal Processing Methods 245

7.2.2 Infant Cry Analysis Methods 247

Appendix A 249

Participant Information Form 249

Appendix B 251

Consent Form 251

Appendix C 252

Analysis of Variance 252

Appendix D 254

Support Vector Machine (SVM) Classifier 254

Appendix E 256

Hidden Markov Model (HMM)...256

E.1 HTK-HMM 258

E.2 The HMM Parameters...258

Appendix F 260

Mel Frequency Cepstral Coefficients (MFCC) 260

References 262

List of Publications...274

Journals...274

Book Chapters 274

Conferences...275

Brief Biodata of the Author...277

Abstract

The present work in this thesis is directed towards understanding the reason of crying of an infant using signal processing approaches. Infant cry analysis and classification is a non-invasive method of analyzing the infant cries and identifying the reason of crying such as pain, hunger, discomfort or presence of any disease. An instrument developed for this purpose may be helpful in bringing up of an infant and preventing the infant from distress because he or she cannot convey his or her needs to the caretakers and thus, improve the quality of life.

For the development of any computer algorithm for the analysis and classification task, development of database is necessary. For infant cry analysis and classification task, if a database is to be created, several factors needs to be considered while designing of the corpus, such as, reason of cry, age of infant, method of cry generation, *etc*. Effect of factors influencing the system behavior developed for infant cry analysis is presented in this thesis. Along with this, ideal characteristics of the infant cry corpus are discussed in this work.

For processing of the infant cry signal, various signal processing challenges associated while using state-of-the-art methods are illustrated in this work. Signal processing methods, namely, Short-time Fourier transform (STFT) analysis, linear prediction (LP) analysis, cepstrum analysis and Teagers energy operator (TEO) analysis are used in this thesis. Along with it, for different pathological cries (such as asthma, meningitis *etc.*) spectrographic analysis is shown.

In this thesis, analysis of different infant cry types is performed using acoustic features such as fundamental frequency (F_0), energy in different frequency bands, unvoicing percentage of cry segments in the cry and duration of cryunits. For understanding the significance of these features in

the cry analysis, *1*-way analysis of variance (ANOVA) is used. Infant cries of various pathological cases are also analyzed using these features and difference in the normal and pathological infant cries are observed and reported in this thesis.

Classification of normal and pathological infant cries is also attempted in this work. For the classification of the normal and pathological infant cries, bispectrum-based features are used and classification accuracy of *81.64 %* is obtained. Performance of bispectrum based features is found to be better than state-of-the-art MFCC features. Noise robustness of the proposed features is also shown.

Classification of two pathologies, namely, asthma and hypoxy ischemic encephalopathy (HIE) are also reported in this thesis. For the classification of these two pathological cry signals, features such as glottal inverse filtering, modulation spectrogram, auditory spectrogram and group delay-based features are used. All these features are found to perform well in classifying these two pathologies. Classification of these pathologies from the normal healthy infants' cries is also attempted in this work. Though the performance of the proposed system is not very good, however, it can help in preventing the infants by giving alarm of presence of HIE disease which can result in motor and physical handicap, if left unattended.

Finally, the work proposed in this work (*i.e.*, analysis of infant cries using prosodic features and classification of normal and pathological infant cries using features based on bispectrum, auditory spectrogram, modulation spectrogram *etc.*) is concluded. Important results from the experiments and analysis of the cries are summarized in the last chapter. This is highly inter-disciplinary field for research. Hence, a lot of scope is there for future research work. Some important research issues are reported in this thesis which can be taken up by researchers in near future.

List of Principal Symbols

$s(\mathrm{n})$	Short segment of speech or infant cry
$\mathrm{S}(m,\omega)$	Short-time Fourier Transform of the signal $s(n)$
$p(n)$	Excitation Source
$h(n)$	Impulse response of vocal tract
$H(z)$	Z-domain system function of vocal tract system
$H(\omega)$	Frequency response of vocal tract system
a_k	LP coefficients
p	LP order
$l(n)$	Lifter in time (quefrency)-domain
$\psi\{\mathrm{x(n)}\}$	TEO profile of signal $x(n)$
E_{1n}	Normalized energy in *0-1* kHz band
E_{2n}	Normalized energy in *1-3* kHz band
E_{3n}	Normalized energy in *3-5* kHz band
E_{4n}	Normalized energy in *5-6* kHz band
$m_n^x(\tau_1,\tau_2,...,\tau_{n-1})$	n^{th} order moment function
$c_n^x(\tau_1,\tau_2,...,\tau_{n-1})$	n^{th} order cumulant function
$B(\omega_1,\omega_2)$	Bispectrum
$C(\omega_1,\omega_2,\omega_3)$	Trispectrum
A	Tensor
U_{I_2}	Unitary matrix in dimension I_2
$X_l(k,i)$	Modulation spectrogram with dimension k and i
$\tau_c(\omega)$	Modified group delay function

List of Acronyms

ANOVA	Analysis of Variance
BRAC	Basic Rest Activity Cycle
CNS	Central Nervous System
EIC	Epoch Interval Contour
ESC	Epoch Strength Contour
FT	Fourier Transform
GA	Gestation Age
HIE	Hypoxy Ischemic Encephalopathy
HMM	Hidden Markov Model
HOSA	Higher-Order Spectral Analysis
HOSVD	Higher-Order Singular Value Decomposition
HT	Healthy Term
IEC	Institutional Ethics Committee
LPC	Linear Prediction Coefficients
MCC	Mathew's Correlation Coefficient
MFCC	Mel Frequency Cepstral Coefficients
OPD	Outdoor Patient Department
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PLPCC	Perceptual Linear Prediction Cepstral Coefficients
RBF	Radial Basis Function
SIDS	Sudden Infant Death Syndrome
STFT	Short -Time Fourier Transform
SVM	Support Vector Machine

List of Tables

Table 2.1. Literature survey of infant cry classification....................................36

Table 3.1. Distribution of infants 48

Table 3.2. Distribution of samples over gender...48

Table 3.3. Detailed distribution of samples over gender for *Corpus I* 48

Table 3.4. Distribution of samples over cry type for *Corpus I*............................49

Table 3.5. Distribution of samples over age for *Corpus I*...................................49

Table 3.6. Distribution of samples over age for *Corpus II*..................................49

Table 3.7. Distribution of samples over pathologies for *Corpus I*......................49

Table 3.8. Distribution of samples in special cases for *Corpus I* 49

Table 3.9. Distribution of cry recordings in different health conditions for *Corpus II* 50

Table 3.10. Distribution of pathological cry samples among several pathologies in *Corpus II* 50

Table 3.11. Number of cry recordings of normal, deaf and asphyxia cries in *Corpus III*...50

Table 3.12. Number of cry samples (*1s* duration) in various health conditions in *Corpus III* 51

Table 3.13. Variation of acoustic features with age on *Corpus I* 53

Table 4.1. Variation of formants with speaker age and gender (formant frequencies are in kHz) (the data is not shown for the same subject)..............70

Table 5.1. Window functions used in the analysis 108

Table 5.2. Presence of different modes in the spectrogram of a cry 123

Table 5.3. Spectral energy features of adult speech 127

Table 5.4. Spectral energy features of infant cries 128

Table 5.5. Classification accuracy (in %) using spectral energy features for the classification of adult speech and infant cries...129

Table 5.6. Statistics of features for all cry types 132

Table 5.7. ANOVA analysis of different cry types 133

Table 5.8. Corpus statistics for infant cry analysis 136

Table 5.9. Distribution of cry samples of newborn infant cries 137

Table 5.10. Mean values of the features for different infant cry types 142

Table 5.11. *p*-values obtained from ANOVA analysis of the newborn infant's cry. 144

Table 5.12. Number of infant cries for different classes 155

Table 5.13. ANOVA analysis of infant cry signal for significance of unvoiced frames 156

Table 5.14. Mean values of unvoiced frames in a cry episode. Adapted from [**152**] 157

Table 5.15. ANOVA analysis of infant cry signal for significance of F_0 157

Table 5.16. Mean values of F_0 in a cry episode 158

Table 5.17. Values of energy features for high risk, respiratory distress and normal infants' cries 165

Table 6.1. Classification accuracy (in %) for 4-fold cross-validation using holdout method 182

Table 6.2. Confusion matrix of classification of normal and pathological cries using bispectrum as a feature set. Adapted from [**171**] 182

Table 6.3. Classification accuracy (in %) with MFCC, LPC, PLP and bispectrum features. Adapted from [**171**] 182

Table 6.4**.** Classification accuracy (in %) with MFCC, LPC, PLP and bispectrum features 184

Table 6.5**.** Confusion matrix of classification of normal and pathological cries using bispectrum as a feature 184

Table 6.6. Comparison of classification performances (in %) of bispectrum features extracted from existing methods under same experimental setup. Adapted from [**172**] 184

Table 6.7. Comparison of classification accuracy (in %) using different feature sets for the classification of normal and pathological infant cries 190

Table 6.8. Classification accuracy (in %) with different FFT bin sizes and lags in bispectrum estimation using the indirect method.......................................191
Table 6.9. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using bispectrum (Direct method) 191
Table 6.10. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using bispectrum (Indirect method) 191
Table 6.11. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using MFCC feature set 191
Table 6.12. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using LPC feature set 191
Table 6.13. Comparison of performance of feature extraction methods from bispectrum (direct method) on classification accuracy (in %).........................193
Table 6.14. Comparison of performances of various feature sets using various statistical parameters 194
Table 6.15. Effect of different noises on classification performance (in %) when indirect method of bispectrum is used 196
Table 6.16. Effect of babble noise on classification performance (in %) on standard features...196
Table 6.17. Effect of babble noise on classification performance (in MCC) on standard features...196
Table 6.18. Effect of car noise on classification performance (in %) on standard features...196
Table 6.19. Effect of car noise on classification performance (in MCC) on standard features...197
Table 6.20. Effect of HF Channel noise on classification performance (in %) on standard features...197
Table 6.21. Effect of HF Channel noise on classification performance (in MCC) on standard features 197
Table 6.22. Effect of white noise on classification performance (in %) on standard features...197

Table 6.23. Effect of white noise on classification performance (in MCC) on standard features........198

Table 6.24. Classification accuracy (in %) with MFCC, LPC, PLP and bispectrum features. Adapted from [**172**]........199

Table 6.25. Performance of the features for classification of normal, asthma and HIE infant's cries........204

Table 6.26. Confusion matrix for the classification of asthma, HIE and normal infant's cries using PLPCC features. Mean classification accuracy is *53.77* %........204

Table 6.27. Confusion matrix for the classification of asthma, HIE and normal infant's cries using LPC features. Mean Classification accuracy is *62.63* %..204

Table 6.28. Confusion matrix for the classification of asthma, HIE and normal infant's cries using MFCC features. Mean classification accuracy is *61.00* % 205

Table 6.29. Confusion matrix for the classification of asthma, HIE and normal infant's cries using bispectrum features. Mean classification accuracy is *54.1* %........205

Table 6.30. Classification accuracy (in %) of asthma and HIE pathological infant cries with modulation spectrogram features applied on SVM classifier with RBF kernel with parameter $\gamma = 0.01$. Adapted from [**173**]........210

Table 6.31. Confusion matrix for classification of asthma and HIE infant cries in *experiment 1* for feature dimension of *100. Adapted from* [**173**]........210

Table 6.32. Confusion matrix for classification of asthma and HIE infant cries in *experiment 2* for feature dimension of *100*........210

Table 6.33. Comparison of modulation spectrogram features (for feature dimension of *100*) with MFCC for classification of asthma and HIE infant cries........211

Table 6.34. Comparison of modulation spectrogram (for feature dimension of *100*) features with MFCC for classification of asthma and HIE infant cries with varying SNR........211

Table 6.35. Classification accuracy (in %) with different features when features are arranged in the form of a vector for a cryunit for the classification of asthma and HIE infant cries....217

Table 6.36. Classification accuracy (in %) with different features when mean of the features is taken for a cryunit using holdout method for the classification of asthma *vs.* HIE infant cries....218

Table 6.37. Classification accuracy (in %) for HIE *vs.* asthma infant cries with different features when mean of the features is taken for a cry using holdout method with feature dimension reduction. Adapted from [**188**]....218

Table 6.38. Distribution of cryunits in training and test datasets for *10*-fold cross-validation experiment....226

Table 6.39. Classification accuracy (in %) in *10*-fold cross-validation experiment. Adapted from [**191**]....227

Table 6.40. Classification accuracy (in %) with MFCC and proposed MODGRD features with different degrees of polynomial kernel function applied to SVM classifier....233

Table 6.41**.** Classification accuracy (in %) with MFCC and proposed MODGRD features with radial basis function (RBF) kernel applied to SVM classifier. Adapted from [**196**]....233

Table 6.42**.** Confusion matrix with proposed feature set using *4th* order polynomial kernel function....233

Table 6.43**.** Confusion matrix with MFCC feature set using *4th* order polynomial kernel function. Adapted from [**196**]....234

Table 6.44. Confusion matrix with proposed feature set using $\gamma = 0.25$ with RBF kernel function....234

Table 6.45**.** Confusion matrix with MFCC feature set using $\gamma = 0.25$ with RBF kernel function....234

Table 6.46. Classification of cryunits of individual infants in the two classes using proposed and MFCC features....234

Table 6.47. Distribution of cryunits in normal and deaf class....238

Table 6.48. Classification accuracy (in %) with different features for the frame size of *30* ms and sampling frequency of *8* kHz. 239
Table 6.49. Classification accuracy (in %) with different features for the frame size of *30* ms and sampling frequency of *16* kHz. 239
Table 6.50. Classification accuracy (in %) with different features for the frame size of *100* ms and sampling frequency of *8* kHz. 239
Table 6.51. Classification accuracy (in %) with different features for the frame size of *100* ms and sampling frequency of *16* kHz 240
Table C. 1 ANOVA Table. After [200]…………………………………………..253
Table D. 1 Various kernel fuctions used in SVM………………..……………..254

List of Figures

Figure 1.1. Flowchart of the system architecture for infant cry analysis and classification. 11

Figure 1.2. Flowchart of the thesis. 13

Figure 2.1. Physioacoustic model of cry production. After [**27**]. 21

Figure 2.2. Organization of CNS in infants. After [**27**]. 23

Figure 3.1. Spectrogram of a *2* months old infant's cry sample (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a). 52

Figure 3.2. Spectrogram of a *6* months old infant's cry (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a). 52

Figure 3.3. Spectrogram of a *12* months old infant's cry sample (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a). 52

Figure 3.4. Variation of acoustic features in infant cry with age (F_s= *12* kHz). Panel I: birth cry, Panel II: Day 1 cry, Panel III: day 4 cry, Panel IV: day 11 cry and Panel V: day 20 cry. 54

Figure 3.5. Infant of *2.1* kg weight of age *4.5* months of age (F_s= *44.1* kHz). (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a). 55

Figure 3.6. Infant of *6.2* kg weight of age *5* months of age (F_s= *44.1* kHz). (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a). 55

Figure 3.7. Spectrogram of an infant cry suffering from laryngomalacia taken from *Corpus II*. 58

Figure 3.8. Cry modes present in the spectrogram of an infant cry suffering from asthma sample taken from *Corpus II*. 59

Figure 3.9. Cry modes present in the spectrogram of an infant suffering from congenital heart disease sample taken from *Corpus II*. 59

Figure 3.10. Spectrogram and cry modes present in the spectrogram of a healthy normal infant's cry sample taken from *Corpus I*. 60

Figure 3.11. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II-*12* samples, Panel III- *20* samples and Panel IV- *25* samples for infant cry signal at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) real cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response superimposed on STFT obtained by liftering (b). 63

Figure 4.1. Short-time Fourier spectrum of a male speech for vowel /aa/ (a) time-domain signal and (b) corresponding short-time Fourier spectrum at *Fs*= *12* kHz. 68

Figure 4.2 Short-time Fourier spectrum of female speech for vowel /*aa*/ (a) time-domain signal of and (b) corresponding short-time Fourier spectrum at *Fs*= *12* kHz. 69

Figure 4.3. Short-time Fourier spectrum of an infant cry (a) time-domain signal and (b) corresponding short-time Fourier spectrum at *Fs*= *12* kHz. 69

Figure 4.4: Short-time Fourier analysis of male (Panel I), Female (Panel II), Children's speech (Panel III) and infant's cry signal (Panel IV). In all the Figures, (a) shows time-domain signal and (b) shows short-time Fourier spectrum of the signal shown in (a). 71

Figure 4.5. All-pole model via Linear Prediction (LP) of the speech signal. After [**119**].. 72

Figure 4.6. Cascade of second order digital resonators for speech generation using LP model.. 73

Figure 4.7. LP analysis of a male speech for vowel /*aa*/ (a) time-domain signal (b) corresponding LP residual and (c) corresponding LP spectrum for LP order of *12* shown by dark line superimposed on STFT 75

Figure 4.8. Effect of LP order *p* in male speech analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures, F_s=*12* kHz. In each subfigure (c), the amplitude is in dB. 77

Figure 4.9. Effect of LP order *p* in female speech analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures: (a) and (b) X- axis is sample index and in subfigure

(c) the X-axis is the frequency in Hz. In all the subfigures, Y-axis is amplitude, F_s=12 kHz. In each subfigure (c), the amplitude is in dB. 78

Figure 4.10 Effect of LP order p in child speech analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures: (a) and (b) X-axis is sample index and in subfigure (c) the X-axis is the frequency in Hz. In all the subfigures, Y-axis is amplitude and F_s=12 kHz. In each subfigure (c), the amplitude is in dB. 79

Figure 4.11. Effect of LP order p in the infant cry signal analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures: (a) and (b) X- axis is samples and in subfigure (c) the X-axis is the frequency in Hz. In all the subfigures, Y-axis is amplitude and F_s=12 kHz. In each subfigure (c), the amplitude is in dB. 80

Figure 4.12. LP order *vs*. LP root mean square (rms) error. 81

Figure 4.13. Transition of formants into F_0 harmonics with increasing order of linear predictor (LP). (a) male speech, (b) female speech, (c) children speech and (d) infant cry signal. 83

Figure 4.14. Cepstrum analysis of a speech segment of female speaker (a) time-domain signal, (b) real cepstrum of (a) and (c) estimated cepstrally smoothed vocal tract frequency response with lifter size *20* samples superimposed on STFT. 85

Figure 4.15. Cepstrum analysis for a male speech at the sampling frequency of *12* kHz with varying lifter sizes. The lifter sizes are: Panel I- *10* samples, Panel II-*20* samples , Panel III- *25* samples and Panel IV- *30* samples. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT. 87

Figure 4.16. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II- *15* samples , Panel III- *20* samples and Panel IV- *25* samples for a female speech at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal,

(b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT...88

Figure 4.17. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II-*15* samples, Panel III- *20* samples and Panel IV- *25* samples for child speech at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT...89

Figure 4.18. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II-*12* samples, Panel III- *20* samples and Panel IV- *25* samples for infant cry signal at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response superimposed on STFT obtained by liftering (b). 90

Figure 4.19. Panel I- Voiced infant cry signal and Panel II- Voiced infant cry signal preceded by a silence. In all subfigures: (a) Time-domain signal of an infant's cry and (b) corresponding TEO profile and (c) LP residual of (a).......91

Figure 4.20. TEO analysis of voiced normal infant cry signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a)......................92

Figure 4.21. TEO analysis of voiced male speech signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a). 92

Figure 4.22. TEO analysis of voiced female speech signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a)......................93

Figure 4.23. TEO analysis of voiced children speech signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a)......................93

Figure 5.1. Formation of cryunits from a cry utterance (a) time-domain cry signal and (b) narrowband spectrogram of (a)...99

Figure 5.2. Wideband spectrogram of (a) normal male speech, (b) normal female speech, (c) normal child speech and (d) infant cry...............................101

Figure 5.3. (a) Time-domain waveform of the cry signal (b) wideband spectrogram and (c) narrowband spectrogram..104

Figure 5.4 Effect of variation in window length on the STFT analysis. (a) time-domain infant cry signal and its narrowband spectrogram with window size of (b) *1.5 ms*, (c) *5 ms*, (d) *10 ms*, (e) *15 ms*, (f) *20 ms*, (g) *30 ms*, (h) *50 ms*, (i) *70 ms* (j) *100 ms*........105

Figure 5.5 Effect of variation in window length on the frequency response characteristics of the window. Window lengths are (a) *1.5* ms, (b) *5* ms (c) *10* ms and (d) *20* ms. 106

Figure 5.6 Effect of selection of window on the spectrogram. (a) Time-domain signal and its corresponding spectrogram using (b) rectangular (c) Hamming (d) Hanning (e) Gaussian (f) Kaiser (g) triangular (h) flat top windows........107

Figure 5.7. Various window functions and their frequency responses. (a) Rectangular window (b) Hamming (c) Hanning (d) Gaussian (e) Kaiser (f) triangular and (g) flattop windows........108

Figure 5.8. Infant cry modes from narrowband spectrogram (a) flat (b) rising (c) falling (d) double harmonic break (e) glottal roll (f) vibrations (g) hyperphonation (h) inspiratory phonation (i) dysphonation and (j) weak vibration. 110

Figure 5.9. (a) Time-domain waveform, (b) corresponding spectrogram and cry modes present in the spectrogram of a normal infant's cry........111

Figure 5.10. (a) Time-domain waveform, (b) corresponding spectrogram and cry modes present in the spectrogram of a normal infant's cry........111

Figure 5.11. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of a neonate's cry. 112

Figure 5.12. (a) Time-domain infant cry waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from laryngomalacia. 113

Figure 5.13. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from asthma........114

Figure 5.14. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from congenital heart disease. 115

Figure 5.15. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from Down's syndrome. 116

Figure 5.16. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and present in the narrowband spectrogram of an infant suffering from brain hemorrhage. 116

Figure 5.17. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from malnutrition. 117

Figure 5.18. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from HIE. 118

Figure 5.19. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from hydrocephalus. 118

Figure 5.20. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of two different infants suffering from meningitis. 120

Figure 5.21. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from respiratory distress. 120

Figure 5.22. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of a deaf infant's cry. 121

Figure 5.23. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from asphyxia..122

Figure 5.24. Effect of various additive noises on narrowband spectrograms of an infant cry at SNR level of (*10 dB*). Spectrogram for (a) clean signal, (b) babble noise, (c) white noise, (d) car noise and (e) HF channel noise. 124

Figure 5.25. Effect of various noises on narrowband spectrograms of an infant cry at SNR level of (*5 dB*). Spectrogram for (a) clean signal, (b) babble noise, (c) white noise, (d) car noise and (e) HF channel noise. 124

Figure 5.26. Effect of various noises on narrowband spectrograms of an infant cry at SNR level of (*-5 dB*). Spectrogram for (a) clean signal, (b) babble noise, (c) white noise, (d) car noise and (e) HF channel noise. 125

Figure 5.27. Scatter plots of features (a) E_{1n} and E_{2n} and (b) E_{1n} and E_{3n} (c) E_{3n} and E_{4n} derived for adult speech and infant cries. In all the subfigures, 'o' corresponds to adult speech and '+' corresponds to infant cry signal...........128

Figure 5.28. Probability density functions (pdf) of (a) E_{1n} and (b) E_{3n} for adult speech and infant cries. 128

Figure 5.29. Estimation of pitch period using autocorrelation method. (a) time-domain infant cry signal, (b) its autocorrelation function and (c) peaks corresponding to pitch period. In all the subfigures, X-axis is sample index and Y-axis corresponds to the amplitude of the signal. 131

Figure 5.30. Box plots of duration features for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all subplots, Y-axis is time in sec. and X-axis represents the class type. **Notations**: N: Normal, P: Pathological, NP: Normal Pain, NH: Normal Hunger, PP: pathological pain, PH: Pathological hunger infant cry. 133

Figure 5.31. Box plots of minimum fundamental frequencies (F_0) feature for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is the frequency in Hz and X-axis represents the class type...134

Figure 5.32. Box plots of maximum fundamental frequencies (F_0) feature for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is the frequency in Hz and X-axis represents the class type....135

Figure 5.33. Box plots of mean fundamental frequencies (F_0) feature for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is the frequency in Hz and X-axis represents the class type....135

Figure 5.34. Modified autocorrelation algorithm for F_0 extraction for Panel I: voiced segments and Panel II: unvoiced frame. In all the subfigures (a) time-domain signal, (b) center clipped signal of (a) and (c) autocorrelation function of (b). 139

Figure 5.35. Comparison of pitch extraction methods (a) autocorrelation method (b) modified autocorrelation method. 139

Figure 5.36. Boxplot for the F_0 features (a) mean F_0 and (b) median F_0....143

Figure 5.37. Barplot of mean values of normalized energy values for different cry types. Y-axis represents the normalized energy of the signal. 144

Figure 5.38. Bar plot of mean values of *E1 (0-2 kHz)* for different types of cries. Y-axis represents the normalized value of feature *E1.* 145

Figure 5.39. Bar plot of mean values of *E2* (2-4 kHz) for different types of cries. Y-axis represents the normalized value of feature *E2.* 145

Figure 5.40. Bar plot of mean values of *E3 (4-6 kHz)* for different types of cries. Y-axis represents the normalized value of feature *E3.* 145

Figure 5.41. Boxplots of normalized (a) *E*, (b) *E1*, (c) *E2* and (d) *E3* for birth and pain cries. 147

Figure 5.42. Boxplots of normalized (a) *E*, (b) *E1*, (c) *E2* and (d) *E3* for birth and normal cries. 147

Figure 5.43. Boxplot for the UV ratio in the infant cries. 148

Figure 5.44. Extraction of pitch values from TEO profile. (a) cry signal and (b) TEO profile of (a). Adapted from [**152**]..152

Figure 5.45. Extraction of pitch values from TEO profile (unvoiced signal). (a) cry signal and (b) TEO profile of (a)...152

Figure 5.46. Spectrogram of infant cry: (a) infant cry signal and (b) spectrogram and fundamental frequencies (F_0) contour extracted from TEO (a), Y-axis represents NFFT bins which corresponds to frequency in Hz). Adapted from [**152**]...153

Figure 5.47. Comparison of fundamental frequencies (F_0) estimation algorithms (a) TEO-based and (b) autocorrelation-based method.................153

Figure 5.48. Comparison of fundamental frequencies (F_0) estimation algorithms (a) TEO-based and (b) modified autocorrelation-based method. 154

Figure 5.49. Boxplot of the unvoicing ratio in infant cries..............................156

Figure 5.50. Spectrogram of newborn birth cry. X-axis is frame index and Y-axis represents FFT bins which corresponds to frequency (*128* bin= *6000* Hz). 159

Figure 5.51. Bar plots of energy in the infants with high risk, RDS and normal health condition...165

Figure 6.1. Distribution of skewness and kurtosis features for (a) normal and Gaussian signals, (b) pathological and Gaussian signals and (c) normal, pathological infant cries and Gaussian signals...170

Figure 6.2. Skewness and kurtosis feature distribution for adult voices........170

Figure 6.3. Bispectrum patterns for Gaussian signal, normal cry signal and pathological cry signal. In all the subfigures, Panel (a) time–domain waveforms, Panel (b) bispectrum using direct method, Panel (c) bispectrum using indirect method and Panel (d) diagonal slice derived from indirect method are shown for Gaussian signal, normal cry signal and pathological cry signal. In subfigures of Panel (a) X-axis is samples and Y-axis is amplitude, in subfigures of Panel (b) and Panel (c), X and Y-axis represents frequencies, ω_1

and ω_2, respectively and in subfigures Panel (d) X-axis is frequency and Y-axis is the amplitude of bispectrum. 172

Figure 6.4. Singular value decomposition of tensor A . After [**170**]. 177

Figure 6.5. Unfolding of tensor A to matrix A_1 and matrix A_2. After [**170**]...178

Figure 6.6. Symmetry regions of bispectrum. 180

Figure 6.7. Classification accuracy (in %) of bispectrum-based features for different feature sizes with *95 %* confidence interval shown with bars. 185

Figure 6.8. (a) Time-domain waveform of infant cry, (b) corresponding narrowband spectrogram, (c) short-time energy of the signal. 186

Figure 6.9. (a) Time-domain waveform and short-time energy of infant cry, (b) cryunit extracted from the signal shown in (a). 187

Figure 6.10. Probability density function (pdf) of additive noises. (a) babble, (b) car, (c) HF channel and (d) white noise. 199

Figure 6.11. (a) Time-domain waveform, (b) corresponding spectrogram of asthma cry sample, (c) time-domain waveform and (d) corresponding spectrograms of HIE cry samples. In all the subfigures, X-axis represents time in sec. 201

Figure 6.12. Modulation spectrogram: (a) HIE (b) asthma. In all subfigures, X-axis denotes modulating frequency in Hz, and Y-axis denotes acoustic frequency in Hz. 207

Figure 6.13. Iterative adaptive inverse filtering (IAIF) method for glottal flow waveform estimation. After [**185**]. 215

Figure 6.14. (a) Time-domain (b) glottal flow waveform of an *asthma* cry segment. 219

Figure 6.15. (a) Time-domain (b) glottal flow waveform of a *HIE* cry segment. 219

Figure 6.16. (a) Time-domain (b) glottal flow waveform of a normal cry segment. 219

Figure 6.17. Block diagram of the auditory (cochlear) model. After [**189**]. 220

Figure 6.18. (a) Cry segment of a HIE cry, (b) its spectrogram and (c) corresponding auditory spectrogram. 223
Figure 6.19. (a) Cry segment of an asthma cry, (b) its spectrogram and (c) corresponding auditory spectrogram. 224
Figure 6.20. Selection of cryunit from infant cry recording. 225
Figure 6.21. Difference in proposed feature and auditory spectrogram (a) cry segment, (b) auditory spectrogram and (c) auditory spectrogram after logarithm. 227
Figure 6.22. Distribution of energy in different frequency components of all cry samples for (a) asthma and (b) HIE. The dotted rectangle indicates the region where energy is higher. 228
Figure 6.23. Energy (mean) distribution across frequencies of asthma and HIE cry samples...228
Figure 6.24. Block diagram for extraction of features from modified group delay function. 232
Figure 6.25. Clustering of feature vectors for the two classes in two-dimensional (*i.e.*, *2-D*) feature space (a) MODGRD and (b) MFCC. 235
Figure 6.26**.** Class separability of MFCC and proposed features. After [**196**]. 236
Figure 6.27. Modified group delay function for normal (Panel A), HIE (Panel B) and asthma (Panel C) infant cries. In all subfigures, (a) corresponds to infant cry signal and (b) corresponds to its group delay function. 236
Figure D.1. Separation of data using SVM classifier with RBF kernel……..…255
Figure F. 1. Block diagram for MFCC feature set extraction…………………260
Figure F. 2. Mel-filterbank used in infant cry analysis in this thesis………260

Chapter 1.

Introduction

1.1 What is an Infant's Cry?

"Infan" is a Latin word which means "speechless" and the word "infancy" came from this word. Infants cannot speak and hence, they use cry signal as their communication language to convey their level of distress. The infant's crying carries only *paralinguistic* content of the speech signal. Crying requires an infant to perform a complicated and sophisticated set of physiological activities that involve the brain, respiratory, motor control and the vocal systems. Crying helps infant's physiology to develop by increasing the pulmonary (lung) capacity [**1**] .

Using crying as communication language, infants convey their needs to their parents. Infant's cry tries to convey to his or her parents any of the following messages, "Mom, I am hungry give me food?" "Dad, I'm getting bored, take me for an outing or play with me". "Something is hurting me in my clothing, please check it". "I'm not feeling well, please help me in getting out of this discomfort". Active parents respond to the call of their infants while some parents leave their infants in distress or alone. Many times, parents observe that their infant is not actually crying because they cannot see tears in his or her eyes. However, actually in infants, tears develop after the age of *4-12* weeks after birth. In such a case, leaving an infant unattended may result in excessive crying behaviour or leaving unattended a medical problem.

To avoid such situations, analysis of infant cry is essential. Efforts have been made in last few decades in this direction of infant cry analysis and classification. Signal processing analysis of infant's cry may help in developing a tool to help parents and infants in communicating the needs of the infants and in the case of presence of any symptom of pathology, help the infant to get medical help without much delay.

In this thesis, efforts are made towards analyzing the infant cry using various signal processing methods. Some methods are proposed for classification of normal *vs.* pathological infant cry. The term *pathological* is defined as the effect caused by some disease [**2**]. Classification of pathologies is also done using various features and statistical evaluation of the results is also presented.

1.2 Motivation

Crying is a symbol of healthy life as soon as an infant comes out of the mother's womb and make his or her presence noticed by the others. This first cry of the infants is the most awaited moment for the parents. After the birth, if an infant does not cry, reason can be many-fold and in that case, the concerned infant has to be investigated by the pediatrician thoroughly. The first cry of a neonate indicates the integration of the organs and neurological system. On one side, where the first cry of the infant gives pleasure to his or her parents, on the other side, it gives a responsibility to the new parents to understand and fulfill the needs of the newcomer to this world, that too without using a formal communication language. In infants, crying is a communication language and we need to understand and decode or decipher it. It conveys not only the need of hunger and discomfort due to wet diaper but also used to express pleasure, pain, colic, loneliness and fussiness by the infants. Failure to understand the reason of crying may leave the baby in danger and at the same time, gives the parents a guilty feeling of not being a good parent. This situation may become adverse if this cycle of infant crying and parent's disability to understand the reason and to sooth the baby continues. It makes the babies "difficult babies" (a baby who cries too much and is difficult to calm) and later on, may result in the infant being ignored and left lonely in his or her family.

To avoid such situation, analysis of the infant cry signal under various reasons of crying is very much required. It can support the parents in making decision about the need of the infant, knowing his or her health status and to get early sign or alert for taking corrective measures as early as possible, if their baby is found to be sick at any point of time. Work in this direction (*i.e*, acoustic and spectrographic analysis) was started in the *1960*s by a Scandinavian team of researchers. Infant cry analysis is a multidisciplinary area of research which requires contributions from pediatrics, neurologists, engineering and linguistics. The infant cry research has helped in decreasing the infant mortality rate and understanding the correlation of cry characteristics with infants' mental and physical development and neural system. In the initial two decades, the cry analysis was done using spectrographic analysis where researchers used spectrograms and defined various distinct cry modes in the spectrogram of the infant cry. The presence of certain cry modes was indicated as the presence of pathology or prone to a pathological condition of infants [**3**].

After successful application of spectrographic analysis for infant's pathology identification, need arises for the development of the computer-based algorithms to detect and classify infants' pathology because spectrographic analysis requires the expertise of the professional to study the spectrograms and identify different cry modes present in it. The automated algorithms have the advantage of minimizing human errors. In this direction, work has been done where researchers have attempted a classification of normal and deaf infants, normal and asphyxiated infants, *etc*. [**4**], [**5**], [**6**]. Apart from the classification of infants from their cry signals, the acoustic signal embedded in infant cry is also used to identify the reason of crying. The reason for crying of an infant can be hunger, pain, to draw the attention of the caretaker, discomfort due to wet diaper or motion or irritability, *etc*. It has been experienced for the years that a mother can identify the reason of crying from the cry sound of babies. Though, it is dependent on other factors such as timing from the last feeding and daily routine of the infant as well. Moreover, it is observed that the parents can identify their infants from their cry as well. All these events indicate the correlation among acoustic features and the reason of crying of the infant. This has motivated the researchers to work on the problem of acoustic analysis of the infant cry and use it, for diagnosis of pathologies. Recently, many applications are developed for infant cry analysis such as “Babypod” in which one can play music to the fetus and fetus can respond to these sounds [**7**], this is done to accelerate the neurological development in newborns. “Crying Bebe” is another application developed by Google play, where infant cry is analyzed for its possible cause [**8**]. However, it cannot distinguish the reason of crying from the pathological perspective.

1.3 Social Relevance of this Research Work

With the urbanization of the country, people are moving to bigger cities in search of jobs. Because of this, most of the people are living in nuclear families. In joint families, people get the experience of the elders of the family in bringing up their children especially during the period of infancy when an infant cannot communicate his or her requirements. Now-a-days, because most of the parents are living in nuclear families, a mother of a newborn, sometimes, cannot identify the reason of crying. A system developed to analyze the cry, may help or assist mothers in such cases. In addition, because the infant cry analysis method is a non-invasive method of cry analysis and it may help in the diagnosis of the pathology to the pediatricians, an application developed for this purpose can be used to boost the confidence while taking a decision by the pediatrician for the diagnosis of particular pathology in infants. Along with this, research in this direction is also of importance to identify the reasons of sudden infant death syndrome (SIDS) and identification and early diagnosis of the infants who are more susceptible to SIDS condition (such as SIDS siblings).

1.4 Applications

If an automated infant cry analyzer can be developed, it can help the society in the following ways:

a. **Identifying the Needs of Infants**: Identifying the features to classify hunger, pain, discomfort cries, can enable the identification of the need of the infant. This can avoid the chances of excessive feeding (most of the time, crying is only related to the feeding requirements by the mother which results in excessive feeding that later on result in vomiting by infant or stomach pain due to gas) and discomfort to the infant and minimize the risk of getting babies irritable. Knowing the reason of the crying correctly may also prevent the chances of wrong parenting which happens in the case of difficult babies.

b. **Development of Medical Support Tool**: As sonography has proved a useful and potential tool in observing the development of the fetus, a cry analyzer developed for identification of the pathologies can be used to support the decision of the doctors. Moreover, in the case of rare diseases, such an instrument or device can give an alarm (or early warning) sign to the doctor so that the infant can be kept under observation and parents can be well informed to see the symptoms of the alarmed pathology. This can avoid the chances of getting late in the cure of the diseases where treatment is available for initial stages of the disease only. It can also help in curing the neurological diseases where the lack of medication results in mental and physical disorders.

c. **Developmental Study of Infants**: In infants, the physiological changes are very fast. These physiological changes in infants are related to the neurological control of the brain. These neurological changes are reflected in the cry of the infants. Thus, the infant cry analysis can be used to study the physical development of the infants.

d. **To Study Language Acquisition in Infants**: The cry patterns of infants also reflect the intonation pattern of the language used in his or her surroundings. The cooing sounds made by the infants are the first step of the infant toward language learning or language acquisition. “How a language is learned or acquired by the infant?”, effect of multilingual atmosphere in child’s language learning, differences in different languages learning pattern can be studied by the infant cry analysis of infants aged above *6* months. In fact, the progress in the development of a model of the speech technology applications such as speech and speaker recognition, language and dialect recognition, speech prosody, speech robotics, *etc.* can be accelerated extensively if we understand how an infant learns or acquires language.

e. **Speech Therapy**: Early diagnosis of the hearing and speech disorders in infants can help the parents and speech therapists to take early initiative towards language learning of such infants.

f. **Study of Prosody**: Since infant’s crying is a non-verbal communication, there is a need to study the messages conveyed by the infant’s crying. The message in the infant crying is hidden in the prosodic content of the cry sound. Thus, it motivates to develop the features which can better convey the information underlying in the infant cry signals. In fact, infant acquires language by exploring speech prosodic uses.

g. **Reduction in Infant Mortality**: The diagnosis of infant pathology on early stages may help in reducing the infant mortality rate.

h. **Study of Speech Robots**: The study of infant cry signal processing is also important in the design of speech robots. To train the robots for the speech content, the infant’s language acquisition study is used and applied to train a robot and vice-a-versa [**9**], [**10**].

These are the some of the applications of the proposed work. Such an application can also help a parent to take care of his or her infant and give him or her healthy life.

1.5 Research Issues in Infant Cry Analysis

Research in infant cry analysis using signal processing methods was started by a team of Scandinavian researchers in *1960*s. In the initial two decades, the cry analysis was mainly using the spectrographic analysis of the infant cry signal. Later on, the cry analysis was tried using automated or semi-automated computer algorithms. In this direction, the work is done towards classification and analysis of infant cry types (such as hunger, pain and pleasure cries), classification of pathological cries where a different set of pathologies were considered by different researchers. However, a little work is done in the field of infant cry analysis and classification of normal *vs.* pathological cries or classification of various pathologies associated with infant cries.

The main challenge in the infant cry analysis and classification is the unavailability of the statistically meaningful database. Collecting a database requires permissions from the hospital authorities and parents as well. Getting cry signals of the pathological infants is furthermore difficult. Thus, getting a statistically significant data or corpus for this task is a challenging task. All the researchers working in this area have their own databases with different sets of cry types, recording conditions, age groups, different pathologies and different weights of infants. Standard database for the task is not available which also restricts the comparison of different research works presented in the literature. As the acoustics of cry signal changes with the cry types such as hunger and pain, it also varies with the age of the infant and the weight of the infant. The effect of all these parameters on the infant cry analysis is demonstrated in this thesis.

All these effects altogether pose a challenge to the researchers to work in this area and contribute towards it. In pathological cry analysis, cry characteristics changes with the severity of the disease. In such cases, long-term follow up of the infant is required which is a difficult task.

The work done in the classification of normal and pathological infant cries is limited to the classification of one type of pathology with normal infant cries. In practical situations, we never know that with which pathology infant is suffering. This creates a need to design a system where a pathological infant cry can be classified from a normal infant cry. For this purpose, we should train the pattern classifier for multiple pathologies and then it needs to be tested for the classification task. In this thesis, an attempt is made to classify normal infant cry *vs.* pathological infant cry where several diseases are considered in pathological class while designing the system. In summary, the challenges associated with infant cry research are as follows:

a. Unavailability of statistically significant database,
b. Getting pathological cry samples of the same disease from same age group of infants is a difficult task, and
c. Classification of pathological cry samples from the normal infant cry samples is a challenging task.

1.6 Contributions in the Thesis

Given several challenges in the infant cry research, an attempt is made towards the analysis and classification of different infant cry types. Following are the area of focus in this doctoral thesis work:

a. **To present the signal processing challenges associated with the infant cry analysis**: Since infant cry signal has higher fundamental frequency (*i.e.,* F_0 is in the range of *350* Hz - *1* kHz) compared to normal adult speakers (*e.g.*, the range of F_0 for male, female and children is *100*

- *150* Hz, *250-350* Hz and *350-500* Hz, respectively), traditional signal processing methods used for adult speech analysis may not work well for the infant cry signal. The application of the speech processing methods to analyze speech of adult speakers to the infant cry signal is reported in this thesis and the need to check the validity of standard methods of signal processing to analyze infant cry signal is being emphasized. High fundamental frequency (F_0) in infants (such as hyperphonic sounds) causes difficulty in identifying formants and their structure from the F_0 due to the sampling of vocal tract spectrum with distantly spaced excitation source harmonics.

b. **Data collection and corpus design for the infant cry analysis**: Analysis of infant cries which includes analysis of normal *vs.* pathological infant cries, hunger *vs.* pain cries, normal *vs.* newborn's cries. Data collection from the human participants (especially, newborn infants) is a difficult and time-consuming task. The method of data collection, ideal characteristics of the corpus, parent consent form, guidelines for data collection recommended by the ethical committee and protection of human rights are explained in the thesis work.

c. **Analysis of infant cry types**: In the present thesis work, analysis of different cry types such as hunger, pain, birth, normal and pathological cries are reported using analysis of variance (ANOVA). Comparison of various cry types is done using features such as pitch (F_0)-based features, duration of the cryunits, spectral energy features and voicing-unvoicing ratio of the cry. The significance of these features in infant cry analysis is reported in the thesis.

d. **Classification of normal and pathological infant cries**: In this case, pathological infant cries are the cries produced by infants who are suffering from some disease (not fever or cold). For feature extraction, bispectrum-based features are proposed. These features are found to outperform the conventional state-of-the-art spectral features, namely, Mel frequency cepstral coefficients (MFCC). Robustness of these features is tested in noisy environments, *i.e.*, under signal degradation conditions.

e. **Classification of the pathological infant cries**: In this case, pathologies considered are asthma and hypoxy ischemic encephalopathy (HIE). Classification of these pathologies is performed using four types of features based on modulation spectrogram, auditory spectrogram, glottal inverse filtering and modified group delay. These features are found to perform excellently in classifying these two pathologies. To quote the statistical significance of the experimental results, *4*-fold cross-validation experiments have been done.

f. **Analysis of high risk infants**: Sudden infant death syndrome (SIDS) is the condition, in which an infant dies all of a sudden without showing any symptoms of sickness. Even after autopsy, the reason of death remains undiagnosed. To study such cases, high risk infants (who may be prone to SIDS) are studied. This class of infants is important to study for identifying the acoustic features which can give an alarming sign to the caretakers so that they can take preventive action at right time and help their child in coming out from the danger of possible death by giving correct treatment on time. This can be proved helpful in reducing the infant mortality rate in India or other countries (wherever this is applicable) and thus, improve the quality of life.

The architecture of the system is explained in the next Section.

1.7 System Architecture

The flowchart for the infant cry classification is shown in Figure 1.1. The cry utterance is pre-processed and then features are extracted for analysis and classification of infant cries. In this work, the features used for the analysis of infant cries are, F_0, duration of the cry, percentage of unvoicing frames in the infant's cry segment and short-time (l^2) energy in different frequency bands. To show the statistical significance of these features in the infant cry analysis literature, analysis of variance (ANOVA) is used.

Classification of normal and pathological infant cries is attempted using features derived from the bispectrum, *i.e.,* higher-order spectrum. Classification performance is shown for both the infant-dependent and infant-independent setup. For dimensionality reduction of the bispectrum features, higher-order singular value decomposition (HOSVD) theorem is used.

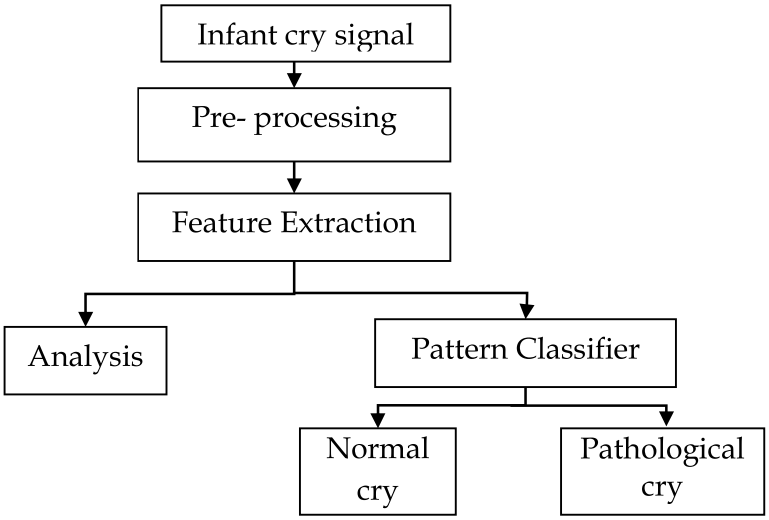


Figure 1.1. Flowchart of the system architecture for infant cry analysis and classification.

Classification of asthma and HIE infant cries is performed using the features derived from the glottal inverse filtering (GIF) of the signal, modulation spectrogram, modified group delay and auditory spectrogram. In all these classification tasks, the performance of the proposed features is compared with the state-of-the-art spectral feature set namely, Mel frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC) and perceptual linear prediction coefficients (PLPC). MFCC is the state-of-the-art method in the infant cry classification task.

Classification accuracy is used as a performance measure. Classification accuracy in percentage is defined for the samples rather than for infants (due to the practical difficulty of getting a large number of infants, especially, for pathological cases, in the database). Another measure used in the work is Matthew's correlation coefficient (MCC). The unbalanced dataset in two classes (namely, normal and pathological infants) are taken care by MCC measure. Along with these, *95* % confidence interval is shown for the results obtained.

For the classification task, support vector machine (SVM) classifier is used in our work mainly with radial basis function (RBF) kernel function. However, in some experiments, other kernel functions are also used. SVM is the standard classifier used for classification experiments. Experimental results are validated using *n*- fold cross-validation.

1.8 Organization of the Thesis

This thesis is organized as follows:

Chapter 1 presents the introduction, motivation and contributions of this doctoral thesis work.

Chapter 2 covers the literature search related to the history of infant cry analysis and methods used for infant cry classification. It gives a brief overview of the literature search about the work done by several researchers in this field. In addition, it presents a brief summary of identified gap area which needs attention and hence, forms a key motivation for the present doctoral thesis work.

Chapter 3 describes the method of data collection and design of the corpus for the experiments. Procedure for data collection (infant cry and other associated metadata) is explained in this chapter of the thesis. In this thesis, three different databases of infant cries are used. Statistics of the data collected is shown in this chapter. Experiences during data collection are also presented.

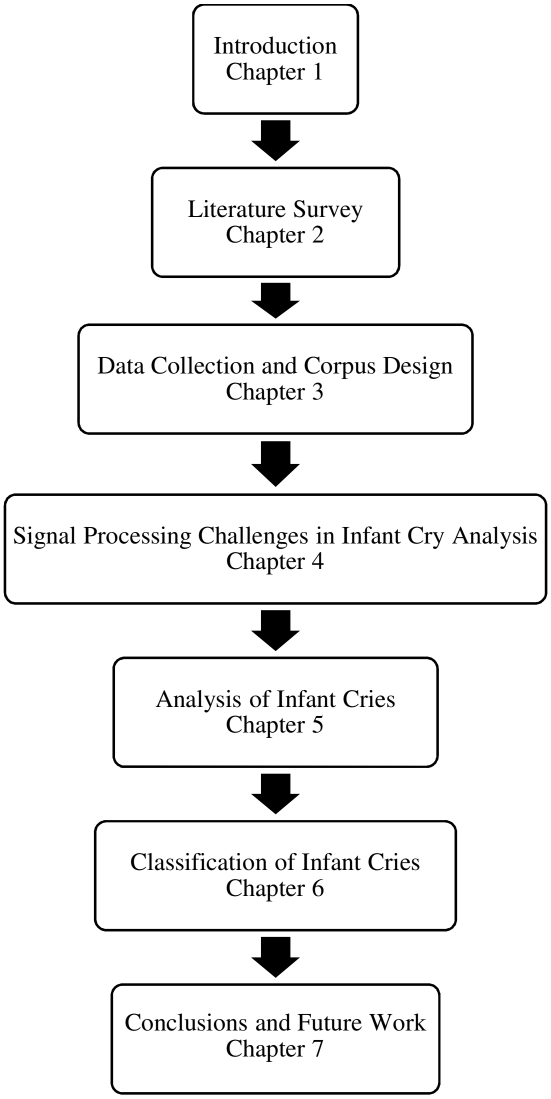

Introduction
Chapter 1
Literature Survey
Chapter 2
Data Collection and Corpus Design
Chapter 3
Signal Processing Challenges in Infant Cry Analysis
Chapter 4
Analysis of Infant Cries
Chapter 5
Classification of Infant Cries
Chapter 6
Conclusions and Future Work
Chapter 7

Figure 1.2. Flowchart of the thesis.

Chapter 4 illustrates the signal processing challenges in infant cry analysis. This chapter shows that how signal processing methods used for adult speech analysis do not work for infant cry signals and thus, pose a difficult signal processing challenge to analyze infant cry. The dependency of infant cry signal analysis methods on various factors is also illustrated in this chapter.

Chapter 5 presents an analysis of infant cries. In this chapter, spectrographic analysis of various pathological infant cries is presented. Analysis for various reasons of crying is also done. In particular, hunger, pain, normal, pathological and newborn cries are analyzed using F_0, duration, the presence of unvoicing in the cry and energy features. The significance of these features is shown using ANOVA analysis and bar plots. Variation of these features with cry types is shown in this chapter.

Chapter 6 explains the classification of normal and pathological infant cries. For the classification task, features derived from the bispectrum are used. Along with this, classification of asthma and HIE infant cries is also attempted. The classification performance is measured using the (%) classification accuracy. Three features are used to classify the two pathologies. Classification of normal infant cries from the cries of HIE and asthma suffering infants are also shown and the results are reported.

Chapter 7 concludes work done in this thesis and also presents limitations of the current research work. It also gives the directions for future research work in infant cry analysis and classification area.

The flowchart of the thesis is shown in Figure 1.2.

1.9 Chapter Summary

In this chapter, motivation and application of using the infant cry analysis and classification, for the research work is illustrated. The challenges in infant cry analysis are also explained which makes the task difficult. The proposed architecture used for the infant cry analysis is given and organization of the remaining thesis is given. In the next chapter, a brief overview of various methods or approaches in the infant cry analysis and classification literature is presented.

Chapter 2.

Evolution of Infant Cry Analysis and Methods

2.1 Introduction

Infant crying has multiple facets for different professionals. Crying is behaviour for a psychologist, which conveys the emotional state of the infant and also informs about the needs of the infant. For a neonate, crying is the way to express his or her physical needs (such as hunger, pain, and wet diaper). From linguistics viewpoint, infant crying is the beginning of vocalization and a step towards acquiring a new language to enable the infant to communicate with the world. For a medical practitioner, newborn crying is an indicator of proper functioning and coordination of different organ and systems and the beginning of a healthy life, whereas shrill cry or no cry at the time of birth is an alarming sign. For an engineer, it is an acoustic event which carries information of the need and physical state of the infant in the form of acoustic descriptors such as melody, loudness, timbre, pitch, intonation, rhythm, *etc.* Hence, it is a multidisciplinary area where professionals from different backgrounds look at the infant cry from a different perspective and it implies that this non-verbal communication is an important part of study in the speech signal processing research. From engineering viewpoint, this area is not matured to the extent of speech and speaker recognition and speech synthesis (which are independent relevant problems in their own right). This is a new field and a lot of work can be explored in the direction of extracting useful information from the infant crying. This thesis is a step towards filling up this gap.

The groundwork for infant cry analysis was laid in the *1960*s by a Scandinavian team of researchers working in Stockholm. Their work resulted in two separate research fields, namely, newborn cry analysis and infant cry analysis [**11**]. Study of infant cry may prove to be effective in the three clinical situations, namely,

a. The cry analysis may be used to support the diagnosis of diseases. In some diseases, the cry has distinct characteristics which draw the attention of the medical practitioner and the parents. Diagnosis through cry analysis which is followed by thorough medical examination may reduce the chances of getting worse due to late diagnosis of infant's health condition in many cases,
b. Early detection of infants at risk can help in reducing infant mortality rate. Infants who are born sick or those who have gone through some trauma or where the siblings have been the victim of SIDS, can be prevented by infant cry analysis system,
c. Infants whose cries are distinct and draw the attention of the listeners are found to be at the risk of SIDS. Such infants can be monitored through the infant cry analysis methods.

Cry is also been studied for the parents' perception of infant cry. It is shown in past that mothers can recognize their infants from their cries. However, recent studies show that fathers are as good as mothers in recognizing their newborn's cries and it depends on the time spent with the infant by parents [**12**]. The perception of a cry from parents is reflected in the parenting depending on how parents perceive the underlying message sent through the cry signal. Cry draws the attention of the parents when the cry is different. Parents response to the infants cry depends on the cry acoustics (F_0, changes in F_0 (*i,e.,* intonation pattern), loudness, duration, rhythm, *etc.*), psychological makeup and living conditions of the parents. Negative response or no response to infant cries may result in child abuse [**13**]. In another study reported in [**14**] shows that the synchrony of arousal between infant and caregiver results in changes in the neurobehavioural mechanisms and the changes in the intensity of arousal are reflected in graded and dynamic acoustic signal. Deviations in the cry signal are noticed by the parents and misunderstanding of these deviations may compromise the infant care and parental effectiveness [**15**]. Moreover, in the same study, it is reported that the infants with abnormal cries should be referred for full neurological evaluation. Infant crying in the early *3* months of the age is a signal of vigor that is evolved to the reestablishment of parental contact [**16**].

In the medical-domain, cry is studied to find out the acoustic features underlying the cry to understand the possible causes of cry and their effects appearant in the cry. Shaken baby syndrome (SBS) is a serious condition resulted from non-accidental head injury with or without impact, resulting from violent shaking. About *25* % of the clinically diagnosed cases die because of it and remaining suffer from lifelong neurological disorders including blindness, learning disabilities and behavioural problems. It has been found that SBS crying peaks at the age of *6* weeks which occurs at *12* weeks in normal infants [**17**]. In the analysis of pain cries of newborns, it is observed that they modulate the supralaryngeal tract considerably following the painful stimulus than in spontaneous cries [**18**]. In invasive pain cry stimulus, the cry produced are rated as urgent by parents and acoustic features such as high fundamental frequency (F_0), longer crying bouts, fewer harmonics and greater variability of the fundamental are found [**19**]. In newborns, dysphonic cries show anger [**20**]. Along with it, newborn cries donot show many differences in the highest and lowest F_0 and there are no differences in the cry acoustics according to gender [**21**]. Gender-specific differences in fundamental frequency and formant frequency patterns appear at the age of *11* [**22**].

2.2 Why Infants Cry?

In the neonate, there is an intrinsic basic sleep / wakefulness cycle as well as a basic rest activity cycle (BRAC) [**11**]. These are controlled by independent biological rhythms. These are controlled by the brain stem with the forebrain mechanism coming into play. In the first one month of life of the neonate, the sleep and wakefulness cycle comes into picture where during the wakefulness period, infant cries for the need of feeding and discomfort due to a wet diaper. During this one month period, the infant gets familiar with the external environment and the sleep/wakefulness cycle is replaced by sleep/alert/wakefulness cycle. As the infant grows, the alert period increases and sleep period reduces. Infant cries not only for feeding requirements, however, for social requirements as well. Increasing alert and decreasing crying is the indicator of balanced exchange between crying and attention. When the balance between the state of alertness and attention is upset, crying results. Crying is a sign of stress that the internal and external requirements are *not* being met. Crying is the part of the regulatory system in which the interplay of behavioural and physiological processes function to maintain homeostatic balance, regulate the duration of alertness and elicit cry when demands are not satisfied. 'Colic' is a special case where infants vigorously and persistently cry (because of pain in the abdomen due to systematic changes in central nervous system (CNS) maturation and muscular hypotenia). It is a gastrointestinal problem [**11**]. Crying occur in response to the stimulation which cannot be accommodated by the system and result in overloading of the system. This is a means of releasing energy and tension.

Cries that are rhythmic are easier for the caretaker to understand and decode. In cry signal, three types of periodicities are found, namely, long periodicities these occur over a long time interval, *i.e.,* some infants cry during the night or in the early morning and shorter periodicities are found in the cry itself, it passes the information of the pitch (or F_0) variations. In this thesis, pitch and fundamental frequency (F_0) are used interchangeably. Cries which are flat in a melody such as in the case of Down's syndrome indicate the abnormal cries and should be immediately taken care of by the parents or guardians. Small periodicities are found in the infant cry, which is the fundamental frequency (F_0) of the cry. The cries with very high or very low F_0 are indicators of the abnormal cries. Even the very high fluctuations in the F_0 pattern are also an indicator of the abnormal health status of the infant.

2.3 The Physiology of Cry and Speech Signal

The adult speech could be made up of phonemes, syllables, words and sentences. The sentences are made of words. In a sentence, punctuations are important to understand the meaning of the sentence. In the adult speech, a speaker can speak *5-7* syllables per second which is called as *speaking rate*. The prosodic clues or content plays a significant role in conveying the message of the sentence. The prosodic content or the suprasegmental information which is given less weight in speech analysis is learnt first by the infant during language acquisition. The suprasegmental features can be observed in the infant cry as well. The two main attributes in learning a new language (syllables) by an infant are curiosity and dynamics of a neural spiking model [**23**]. In newborn infants and the adult speech intonation patterns are similar in following way [**24**]:

a. The duration of the expiratory phase is longer than the inspiratory phase. In infants, the ratio of expiratory phase to inspiratory phase is *2:1.* However, it is *5:1* in adults.
b. Prior to the onset of the phonation, the alveolar pressure rises and at the end of the phonation, it decreases rapidly. Thus, the alveolar air pressure goes through a transition of positive air pressure to negative air pressure during inspiration.

However, the F_0 contour tends to be almost same in the cry also during a non-terminal portion of the cry, whereas in adult speech it varies significantly with respect to time. There is one major difference in the intonation patterns of the adults and newborn infants. Newborn infants cannot regulate the subglottal air pressure by a hold-back of intercoastal muscles. The hold-back mechanism is controlled by the rib cage shape. The ribs are angled downward and outward from the spine in human after the age of *3* months. In infants, the ribs are angled vertical to the spine and thus, infants cannot regulate the subglottal air pressure by working against the air pressure generated by the lungs. As a result, during first three months of the life, the human infants produce phonation of cry of very short duration. The air pressure which is built up prior to the phonation is *2 cm* H_2O in infants as opposed to *8-10 cm* H_2O in adults.

Infants learn to speak through imitation of the adult voices. In order to learn a language, infants produce babbling sounds. Crying and babbling are two different phenomenons. Crying is regarded as a direct connection to adults crying, however, babbling shows structural similarity to the language [**25**]. It has been shown that the fundamental frequency (F_0) contour for a word is similar to the transition of F_0 values in adults and infants. Along with this, infants try to imitate the formant frequency patterns of the adults, however, they cannot imitate the absolute values of the formants because the length of the vocal tract is smaller than the adults.

2.4 Physioacoustic Model of Cry Production

The model of cry production was proposed by Golub in *1979* [**26**]. The model is shown in Figure 2.1. According to the cry production model, the model has four parts, namely, [**27**]:

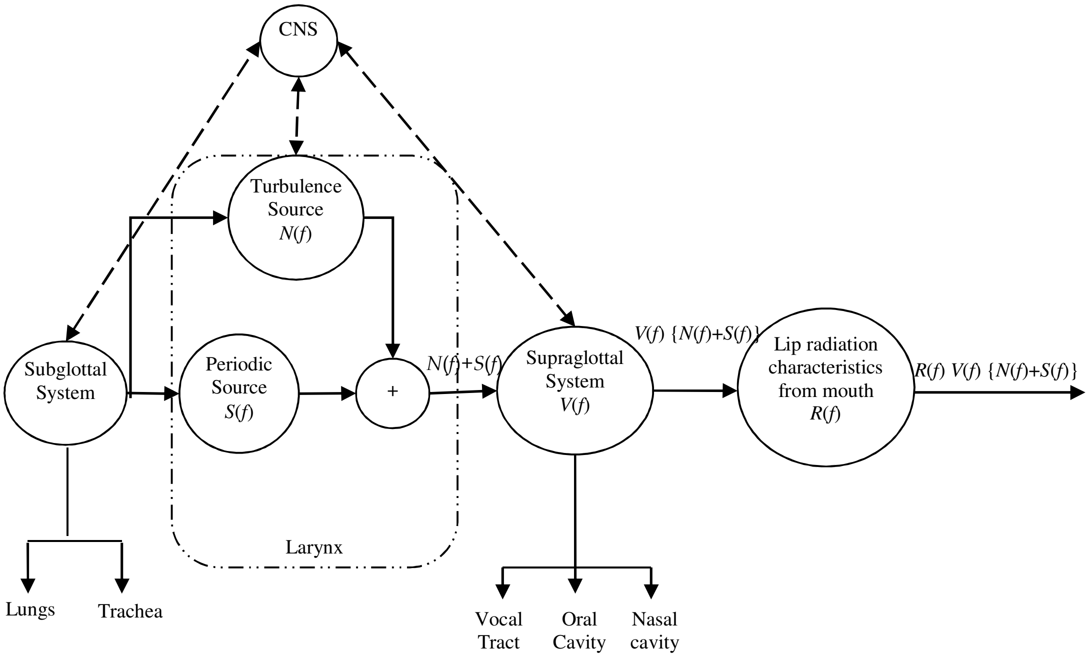


Figure 2.1. Physioacoustic model of cry production. After [**27**].

a. The first part is the subglottal system. The subglottal system or the respiratory system is responsible for developing the pressure $P_s(t)$ below the glottis necessary for generating vocal fold vibrations.
b. The second part is the vocal excitation source, located at the larynx. The source may be either a periodic source $P(f)$ or a turbulent noise source $N(f)$. The periodic source is due to the vibrations of the vocal folds and the turbulence noise source is due to turbulence created by forcing the air through a small opening left by the incomplete closure of vocal folds.
c. The third source is the vocal and nasal tract located above the larynx. This part of cry production acts as a linear time-invariant (LTI) filter and denoted by the transfer function $v(n)$ or frequency response $V(f)$. The response of the transfer function depends on the shape and length of the nasal and vocal tracts and the non-linear coupling between them.

d. The fourth part is the radiation characteristics (also called as lip radiation) $R(f)$ that describes the filtering (typically highpass filtering) of the sound between the mouth of the infant and the microphone.

The acoustical mathematical model which is used for adult speech production can be applied to the infant cry production model and the cry produced can be denoted in mathematical model as follows:

$$X(f)=R(f).V(f).[N(f)+S(f)]. \qquad (2.1)$$

where $X(f)$ is speech spectrum at lips, $R(f)$, $V(f)$ and $[N(f)+S(f)]$ represents spectrum characteristics of lip radiation, supraglottal system and additive turbulence plus periodic source, respectively. From the spectrograms of the cry signal various cry modes are defined, such as, phonation, hyperphonation and dysphonation, *etc.* In phonation, vocal folds of infant vibrate at the F_0 of *250-700* Hz. Hyperphonation results from the falsetto-like vibrations of the vocal folds with F_0 above *1* kHz. In this mode, probably thin portion of the vocal ligament is involved. Dysphonation is due to the turbulent noise at the vocal folds. It contains periodic and aperiodic both vibrations of the vocal folds. All the three modes mentioned above results in the expiratory phase only. Other cry modes from spectrograms are discussed in detail in Section 5.3.4

The filtering (or spectral coloring) of the excitation source by the vocal tract system introduces spectral peaks in the sound output. These peaks are called the formants or formant frequencies. The position of these peaks in the sound output depends on the length of the vocal tract. If the vocal tract is assumed to be of uniform cross-section and the length of the vocal tract is *8* cm (*i.e.*, the length of the vocal tract is *8 cm* in the newborn), then the first two formants, namely F_1 and F_2 will occur at *1100* Hz and *3300* Hz, respectively. If there is substantial velopharyngeal opening, then there will be an additional spectral peak that will occur in the range of *2-3* kHz.

A cry sequence in an infant consists of series of small expiratory cries separated by a brief interval (inspiratory duration). The duration of the expiratory cycle depends on the maturity of the respiratory system. For example, in newborn, the respiratory system is not matured and CNS has poor control on its muscle, which in turn results in shrill or weak cries.

2.5 Organization of Central Nervous System (CNS) in Infants

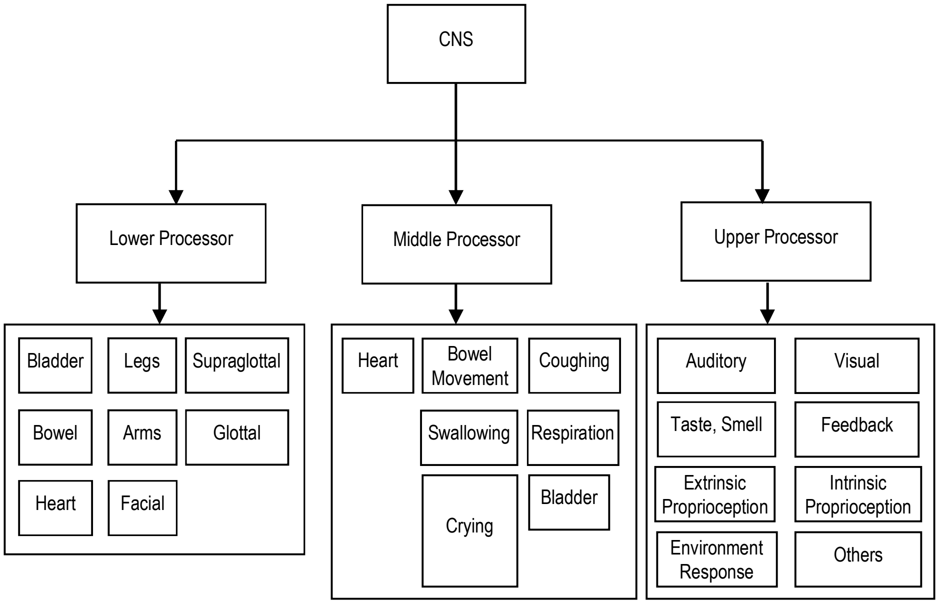


Figure 2.2. Organization of CNS in infants. After [**27**].

Figure 2.2 shows the representation of the overall central nervous system (CNS) in the infants [**27**]. The CNS controls the muscles in three different levels called upper processor, middle processor and lower processor. The upper processor is controlled by the complex internal and external stimulations and their feedback, which result in control of the muscle group for a particular action, that result in choosing and modulating the state of the infant. The middle processor controls the less complex operations in the infant such as swallowing, bowel movement, coughing, respiratory movements, *etc.* [**27**]. In the neonate, the upper processor is not mature and this results in reflex-like cries in the neonatal period. In an infant, following the cry stimulus, upper processor triggers the middle and lower processors which control the relevant muscle group for the control and initiation of the cry. The three mentioned muscle groups important for the cry production are controlled independently [**27**]. The model of the cry production provides a clue for the selection of the features for the acoustic analysis of the cry, those can be later related to the anatomy and physiology of the cry production mechanism [**27**].

2.6 Factors Affecting Crying Behaviour

In human infants, the auditory system matures very quickly. In *15* weeks of gestation age (GA), structural parts of the cochlea and middle ear are formed. Around *20* weeks of GA, middle ear and cochlea become anatomically functional, at *25* weeks of GA, the auditory system becomes functional. At *25-26* weeks of GA, loud noise in utero produces changes in autonomic function (such as, heart rate, blood pressure, respiratory pattern, *etc.*). In as early as *32* weeks of GA, learning occurs, *i.e.*, in utero, fetus recognizes mother's voice, music and common sounds in the environment. In utero, fetus when exposed to low frequency sounds (below C, *i.e.*, caesarean section which is the lower part of the abdomen), may result in damage or even destroy of hair cells (and hence, headphones should not be applied directly on the abdomen during pregnancy). A fetus exposed to intense low frequency sounds may suffer from language delay of two months. [**28**]. Because of this vocal learning in the uterus, an infant is able to identify his or her mother or even respond to music from outside world (such a mechanism is exploited in Babypod product) [**7**] .

A neonatal cry is a coordinated event of three physiological events, namely, the periodicity of CNS regulation (infant moves from deepest sleep to maximal awakeness state), co-ordination of laryngeal muscles with respiratory cycle and nutritional state of the infant (infant state changed from continuous umbilical feeding to discontinuous oral input). An infant cry depends on several other factors such as genetic factors, growth factors and toxic influences [**3**] .

The effect of genetic abnormalities is evident from the cries of infants suffering from trisomy *13-15*, cri du chat and Down's syndrome. In trisomy *13-15* syndrome, the phonation has abrupt onset of phonation, unsteady F_0 patterns and marked a drop of pitch (or F_0) at the end of the phonation [**29**]. Cri-du chat syndrome affected infants cries have a high fundamental frequency (F_0), the absence of vocal fry and flat melody infants with trisomy *13* or *18* show low-pitched cries [**3**].

In the study reported by Michelson [**30**], it was found that the cries of normal infants and cries of infants low to birth (low birth weight normal infants) are similar. Both of these cries are distinct from the premature infant's cries. The more immature (lower age) the infant, the higher is the maximum pitch. Cries of infants with CNS disorders are quite different from the normal infant cries. In addition, birth cries of infants of heroine addicted mothers were found to be high-pitched [**31**].

2.7 Infant Cry Analysis Techniques

The infant cry has been studied for its significance in the developmental perception of the child, integration of anatomy (such as maturation of CNS, coordination among several organs) and for the clinical perspective of diagnosis of unhealthy infants. The cry has been analyzed by the following methods.

2.7.1 Auditory Analysis

It has been observed that mother can recognize the cry of her infant by listening to it. It is also shown by Wasz Hockert *et. al*. that infant vocalizations for different reasons such as hunger, pain, birth and pleasure can be identified auditorily after training [**32**]. It has been found that the pain cries of healthy infants are distinct from the pain cries of infants suffering from some pathology such as neonatal asphyxia, neonatal brain damage, neonatal hyperbilirubinemia and Down's syndrome [**27**]. Though cries can be identified auditorily, the correctness of the decision depends on the training of the person and experience. However, it is not a reliable method and provides a fraction of the information.

2.7.2 Time-Domain Analysis

Time-domain analysis method uses strip chart recorders or direct writing oscillographs to record the infant cry. The durational features are used for the classification of cries. It has been observed that the cries of pathological infants (such as those suffering from a brain hemorrhage) requires higher stimulation to generate the cry of the same duration and the latency for the pathological cries is more than the normal infants. The mean latency time is *1.2- 1.6 sec.* in normal infants whereas it is *2.6 sec* in infants with brain damage. However, the latency time depends on the wakeful/ sleeping state of the infant. The duration of the pain cries in normal infants varies from *2.6-5.2 sec.* However, the duration of phonation is small in sick infants. Michelson *et. al.* have reported that the mean duration of the cries is *1.7* sec in infants with meningitis [**3**]. This is also a useful method to investigate the developmental changes in the infants. As the infant grows, the duration of cries becomes longer. This technique has the advantage of being easy, however, there are losses due to pen inertia, paper speed and human measurement errors. Similar to auditory analysis, this method also imparts fractional information. Features used in the time-domain analysis are as follows:

Latency Period: The time between the pain stimuli applied to the infant and the onset of infant cry sound.

Duration: This feature is measured from the onset of the infant cry to the end of the signal and consists of the total vocalization occurring during a single expiration or inspiration.

Second pause: The time interval between the end of the first cry signal and the following inspiration.

2.7.3 Frequency-Domain Analysis

These methods use a bank of bandpass filters to find the strength of the signal in various frequency bands called as *subbands*. These are used to impart features related to frequency-domain of the signal. These devices give information about the relative magnitude of the various frequency ranges, however, the obtained information alone is of limited value. Features used in the frequency-domain analysis are as follows:

Maximum pitch: The highest measurable point of the fundamental frequency (F_0) seen on the spectrogram.

Minimum pitch: The lowest measurable point in the F_0 contour seen on the spectrogram.

Pitch of shift: Frequency after a rapid increase in the F_0 seen on the spectrogram. Detailed analysis of such F_0 variations for normal and pathological infant cry is presented in Chapter 5 of this thesis.

2.7.4 Spectrographic Analysis

The spectrogram is the visual representation of the signal which represents the distribution of energy in both time and frequency planes. Over the last *20* years, most of the studies in the infant cry analysis were based on the sound spectrographic analysis method. This method utilizes the spectrogram of the cry to derive durational and frequency-related features. These features are then tested for their significance in the analysis of a particular cry type analysis and classification. The durational features used were duration, latency and pauses between the small cryunits. The frequency-related features are maximum pitch, mean pitch, minimum pitch, glottal roll, melody type, gliding, biphonation, *etc*. Spectrographic analysis has proved effective in the diagnosis of the pathologies. Some of the results reported through spectrographic studies are as follows [**3**] :

a. **Cri-du-chat**: In cries of infants with cri-du-chat syndrome, the melody (*i.e.*, prosodic) pattern was found flat and the pitch (or F_0) was found to be in the range of *600-1000* Hz.
b. **Down' s syndrome**: In a study reported on *0-8* months old infants, the vocalization was found long and the mean minimum pitch was *270* Hz and mean maximum pitch was observed at *510* Hz.
c. **Congenital hypothyroidism**: Spectrograms of the *40* cries of *4* infants observed maximum pitch of *470* Hz and mean minimum pitch of *270* Hz. In the spectrogram, glottal roll was observed to occur frequenctly.
d. **Infant's with cleft palate**: There were no significant differences in the fundamental frequency (F_0) with respect to the normal infants. Mean maximum and minimum pitch were *710* Hz and *360* Hz, respectively. Biphonation was not seen in the spectrogram.
e. **Neonatal hyperbilirubinemia**: Maximum and minimum both pitches are high (*2120* Hz and *960* Hz, respectively), biphonation and furcation (*i.e.*, split in F_0) were common in the spectrogram.
f. **Hypoglycemia**: In the spectrogram, biphonation is common in spectrograms. The F_0 is also higher than the normal infants (*1590* Hz compared to *710* Hz).
g. **Asphyxia**: Newborns with central and peripheral asphyxia were studied by the Michelson [**30**]. The study reported that the infants suffering from peripheral asphyxia (*i.e.*, respiratory distress) had a mean maximum pitch of *1000* Hz and in infants with central asphyxia (which is neurological symptoms) had a mean maximum pitch of *1460* Hz. In premature infants, the mean maximum pitch with shift was found to be *1950* Hz in infants with central asphyxia and *1610* Hz in infants with peripheral asphyxia. In normal premature, this parameter

was found to be *1520* Hz. Biphonation was observed in infant cries of asphyxiated infants.

h. **Meningitis**: In the study performed on *14* infants of *0-6* months of age suffering from bacterial meningitis, it is found that the mean maximum pitch is *750* Hz and the mean minimum pitch is *560* Hz. Biphonation is very common in the cries (*49 %*) and glide also occurred frequently in the cries [**33**].
i. **Herpes simplex virus encephalitis**: Noise concentration was found at *2* kHz to *3* kHz region. In the cries of suffering infants, biphonation and glide are common compared to the normal infants.
j. **Hydrocephalus**: The mean maximum pitch shift was *750* Hz and mean minimum pitch is *430* Hz as reported by Michelson *et. al*. [**34**]. The melody is found to be *flat*. Biphonation and glide is also common in this disorder.
k. **Malnutrition**: In infants suffering from kwashiorkor (a form of malnutrition caused by a lack of protein in the diet), the cry characteristics did not differ from the normal infant cries. In marasmus (a form of severe malnutrition) infants, suffering from brain dysfunction showed high mean maximum pitch of *1340* Hz [**35**].

Hence, it is observed that when an infant gets sick, the characteristics of the cry changes from normal to abnormal. In the diseases affecting CNS, the cry characteristics, *e.g.*, the pitch-related features changes, biphonation become common in the spectrogram and the melody shape also changes. The cry analysis can be proved important in the clinical analysis of cry.

2.8 Recent Trends in Infant Cry Analysis

Now-a-days, computer algorithms are used to analyze the infant cry signals, which allow for quick interpretation of results and development of infant cry analysis tools. In the recent developments in this area, work has been done towards infant cry analysis, classification of normal infant cries from sick infants, development of signal processing methods for infant cry analysis and identification of the cry types. Identification of infant from his or her cry is reported in [**36**], [**37**] where the use of MFCC is proposed for this work.

Pioneering work in the infant cry analysis is done by Xie *et. al.* where they defined ten distinct cry modes for automatic classification of the infant's cries [**38**], [**39**], [**40**]. The ten cry modes are flat, falling, rising, vibrations, weak vibrations, glottal roll, double harmonic breaks, dysphonation, inspiratory phonation and hyperphonation (These cry modes are expoited for infant cry analysis in Chapter 5 of this thesis). Using these ten cry modes, a parameter, namely, *H*-value is obtained and this *H*-value is found to be in correlation with the parents rating of the infant's level of distress (LOD). In their work, only newborn infant cries were considered. These infants were healthy infants and the distress is corresponding to the pain of the infants which is due to the vaccination. Pathological cases are not considered in this work and the age group is also restricted to the approximately *39* weeks of GA. In the work of Xie *et. al.*, the cry modes are identified manually, which was automated in the work done by Black [**41**]. In his master's thesis, to estimate the level of distress, *H*-value is estimated. The estimation requires the identification of ten cry modes specified by the Xie *et. al*. [**40**]. Here the *H*-score is defined as

$$H - value = \frac{number\,of\;H\;type\,sequences}{Total\,number\,of\;voiced\;sequences},$$

where the *H*-type cry words are specified as trailing, double harmonic breaks, dysphonation, hyperphonation and inspiration. The cry modes are identified automatically using F_0 and energy features. The cry modes are applied to hidden Markov model (HMM) classifier and accuracy is estimated and compared with parent's perception [**41**] . Again, only normal infant cries were considered in this work, the task of pathological cry classification is not attempted. In Chapter 5, the cry modes suggested by Xie *et. al.* are used to analyze different cries of normal *vs.* pathological infants.

Another interesting work on infant cry analysis shows the effect of delayed auditory feedback on infant crying [**42**]. It has been found that this effect is not consistent for all the ages. In some cases, pitch (F_0) rises, however, in some cases, it falls. Another work shows that excitability in infants is observed through higher duration, F_0 , F_1 and variability in F_1. However, high latency, dysphonation, low utterance duration and low amplitude correspond to the depression in infants. A cry with low duration of utterance, low number of utterances, lower amplitude and higher inspiratory period, short duration utterances, dysphonation and high variability in amplitude represents poor respiratory control. Poor control of vocal tract is associated with higher values of F_0, F_1, F_2, hyperphonation, cry mode changes, variability in F_0 and high variability in F_1 and F_2 [**43**]. Eventhough, these findings are useful in the characterization of infant's health from his or her cry, it was not used to classify the cry signals.

Some researchers have worked in the analysis of first cry or birth cry of the infants. Most of the work in this direction is by the medical practitioners and researchers. In [**44**], authors have used larynx of two newborns (dead) to generate sounds by applying air pressure. Findings show that the role of the larynx is same as excised organ, free of neurologic control. Their role in the first cry is not to vibrate by themselves, however, to generate aerodynamic perturbations generating supraglottic vibrations. Neurological control and regulation is absent in the first cry. Complex interactions are responsible for the nonlinear phenomenon found in the first cry. After the birth, the health of the newborn is rated on the *apgar scale*. Higher apgar count (*maximum 10*) indicates healthy baby and lower apgar counts (< *5*) indicates poor health state of the newborn. In [**45**], authors have proposed an algorithm to automate the computation of apgar scores from the cry signal. In their work, they related the apgar score with the eigenvalues of the principal component analysis (PCA) components for *10* Mel frequency cepstral coefficients (MFCC) coefficients. PCA on MFCC shows that high apgar score cries have high eigenvalues and low eigenvalues are found for low apgar scores. The second, third and fourth coefficients of the MFCC are found to be useful for cry analysis and this result can be used to design an automated algorithm for computing apgar scores.

Attempts have also been made for giving new signal processing techniques for infant cry signals. In this direction, estimation of fundamental frequency (F_0) for infant cry signal is proposed in [**46**], [**47**]. In the case of adult speech, Hiroya Fujisaki was the first to use autocorrelation function to estimate F_0 [**48**]. Cross-correlation-based method is proposed for estimation of infant cry frequency measurement [**46**], [**47**]. In this method, cross-correlation of a frame is found with its nearby frame for a lag τ. Peaks of the cross-correlogram give F_0 of the infant cry. The algorithm works well for adult speech as well and is able to work for double harmonic breaks. In this thesis, statistical validation of the algorithm is not given. Infant *glottogram* is also not available as the ground truth. Only cross-correlograms are shown to validate the theorem. Another method, which was used for F_0 estimation, is average magnitude difference function (AMDF) along with simple inverse filter tracking [**49**]. Along with F_0, *the* autocorrelation function is used to find of voicing or unvoicing segments in the cry. Formants were also used for normal and pathological infant cry analysis. However, statistical analysis is not done. The correctness of proposed cross-correlation-based F_0 theorem (peaks of cross correlation function are used to find pitch period) is not proved (either using spectrogram or using glottogram). In their work, one normal infant cry, one premature infant cry and one pathological infant cry (suffering from phenylketonuria (PKU)) is considered. Results show that more is the deviation from the normal health condition, more irregularities are seen in F_0 contour and the first three formants (*i.e.*, F_1-F_3) set to the lower values. Researchers have proposed a method to find the cry modes in the cry. The method is known as a *5*-line method in which the range of F_0 considered is *200-1000* Hz. Melody shapes are defined using combinations of rising (*+1*), falling (*-1*) and flat (*0*) patterns of F_0 contour. With these combinations, *77* different sequences are possible, however, it is observed that *20* of these melodies include *95* % of the melodies [**50**], [**51**]. The correctness of the proposed method is shown in their work. However, these are not used to classify any cry signals. Cryunit segmentation using short-time Fourier transform (STFT) energy-based algorithm is proposed for automatic segmentation of cryunits [**52**]. In the proposed work, *25* % of maximum energy is taken as the threshold for cryunit selection. Validation is done through a manual *vs.* automatic number of cryunits detected. This method is used by other researchers in the infant cryunit segmentation because it is simple and accurate. Here, accuracy is measured only in terms of the number of cryunits present in the cry, not on the basis of the initial and end timings of the cryunits.

Research has been done in the area of classifying normal infant cries from the pathological infant cries. Most of the work is done towards classifying normal infant cries from the cries of deaf infant or infant with a hearing disorder. MFCC has been used as a feature set for the classification task with different classifier [**53**], [**54**], [**55**]. Another feature used for the normal and deaf babies classification is short-time Fourier transform (STFT) features with time-delay neural network (TDNN), general regression neural networks (GRNN) and multi-layer perceptron (MLP) [**56**]. Analysis of normal and deaf infant cries show that the ratio of the dominant frequency to the fundamental frequency (F_0) is *3* in normal infants while this ratio is 2 in deaf infants [**57**]. On the other hand, results reported in [**58**] shows that the number of cry segments, specific segment length, average pause length and average segment length parameters are not specific to hearing or deafness nor these are good for gender classification. In [**54**], [**55**], [**59**] three-class classification is performed for classification of normal, deaf and asphyxiated infants using features such as MFCC and wavelet-based features. Classification of normal and asphyxia is also attempted using MFCC features in [**5**].

Other work on pathology classification is classification of normal and pathological infant cries where pathologies considered are tetralogy of Fallot, respiratory distress syndrome, intrauterine growth restriction **(**IUGR), hyperbilirubinemia as reported in [**4**], [**60**]. There authors have given an analysis of normal and pathological cries using several features such as fundamental frequency (F_0), phonation, hyperphonation, dysphonation, the number of changes in cry modes and unvoiced sounds. ANOVA analysis is used to show the significance of these features and clear differences has been found among the considered pathologies based on these features. Classification is reported using MFCC feature set. Normal infant cries and cries of infants with cleft palate are also classified using MFCC feature set using HMM [**61**]. Same experimental setup is also used to classify infants with normal health condition with infants suffering from respiratory distress. In [**62**], the work is done to classify normal and infants with cleft palate, preterm infants and sick infants (such as cri-du-chat and down's syndrome) using MFCC, linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC) and fundamental frequency features using HMM. MFCC feature set has also been used to analyze cries of infants suffering from hypothyroidism [**63**]. In the analysis of cries of infants with hypothyroidism, it has been observed that *0*th coefficient of MFCC is not useful and can be ignored in the analysis. Analysis of infant cries of autism disorder is reported in [**64**]. It was found that the pain cry in high risk infants has higher and variable F_0 compared to low risk group.

In last few decades, attempts have been made to classify and analyze different cry types. The cry types defined by several researchers for infants are hunger, pain, pleasure, discomfort, fear, anger and birth cry. Classification of fear, anger and pain cries using MFCC features is reported in [**65**], [**66**]. Hunger *vs.* no hunger and pain *vs.* no pain cries are classified using MFCC feature set with support vector machine (SVM) classifier and NN ensembles [**67**]. Analysis of pain and manipulation cries (cry during changing cloths) is performed using F_0 and F_1- F_3 [**68**]. It was observed that the F_0 and F_1 features are correlated to each other while both work independently for context identification (reason of crying which is pain or changing cloths). F_0 and F_1 contribute significantly to context identification. F_2 and F_3 are not useful for discrimination of pain and manipulation cries. In another study, hunger, pain, and wet diaper cries are classified using the MFCC, short-time energy and pause duration features. Score-level fusion of these features resulted in classification accuracy of *80.56 %* [**69**]. The same cry classification is also performed using epoch and spectral features (MFCC features provide vocal tract system information). Unsupervised Gaussian mixture model (GMM) approach is used for classification [**70**]. In another work by the same authors, epoch contour slope of epoch interval contour (EIC) and epoch strength contour (ESC), epoch sharpness features are used for the same task [**71**], [**72**]. Some of the important results of classification of infant cry types in the literature are shown in Table 2.1.

Table 2.1. Literature survey of infant cry classification

S. No.	First Author	Problem	No. of infants	Features	Experimental Results
1.	Avinash Singh et. al. **[70]**	Hunger, pain & wet diaper	*120*	MFCC, EIC	GMM 78.96 %
2.	Sadra E. Barajas **[67]**	Hunger and pain	No. of infants not mentioned, howevr, *209* pain samples and *1418* no pain and another corpus with *759* hunger and *868* no hunger	MFCC and its *16* Principal components	Pain no pain 96.41% Hunger no hunger 87.61 % using NN and SVM ensemble
3.	M. Petroni **[65]**	Anger, fear and pain	*16* infants *195* cry episodes used out of *230*	MFCC	Art. NN: FFNN best 79.4 %
4.	Hariharan **[56]**	Normal and deaf	*6* deaf and *5* normal (baby chillanto) *1* sec segments	STFT	*93.9* % using GRNN (General regression NN)
5.	Jose Orozco Garcia **[53]**	Normal and deaf	*78* recordings of *31* infants *157* normal and *879* deaf cryunits	MFCC with PCA (*50*)	*97.43* %
6.	José Orozco **[73]**	Normal *vs.* deaf	*31* total infants *157* normal and *789* pathological cryunits of *1* sec	LP (*16*) with PCA	*86.2* % (*10*- fold) Scaled conjugate gradient NN
7.	Rosales Perez **[74]**	1. Asphyxia *vs.* Normal 2. Deaf *vs.* normal 3. Hunger *vs.* pain	Asphyxia: *340* Deaf :*879* Normal : *507* Hunger *350* Pain: *192* Baby chillanto	MFCC (*16* coefficients, *50* msec frame size)	1. *90.68* % 2. *99.42* % 3. *97.96* % Using genetic selection fuzzy model (GSFM)
8.	Hariharan **[75]**	Normal *vs.* Asphyxia	Baby chillanto	STFT	*99* % with Prob. NN
9.	M. Hariharan **[76]**	Normal *vs.* asphyxia	Baby chillanto	Wavelet db *5* frequency band-level	*99* %
10.	Sahak **[77]**	Normal *vs.* Asphyxia	University of Milano–Bicocca *316* healthy *284* asphyxia	MFCC	*95.86* % SVM RBF
11.	Galaviz Orion **[78]**	Normal, deaf, asphyxia	*1049* normal *879* deaf *340* asphyxia	MFCC and LPC	*96.79* % with MFCC *100* ms *86.75* % with LPC
12.	Orion F. Reyes-Galaviz **[79]**	Normal, deaf *vs.* asphyxia	*1049* samples of normal *340* asphyxia *879* deaf	LPC and MFCC	LPC : *76.06* % MFCC: *86.06* % using input delay NN (*3* class)

13.	Orion Fausto [**55**]	Normal and pathological (deaf and asphyxia)	No. of infants not clear *0.4 s* cryunits *334* normal and *549* pathological	MFCC with PCA (*50* coeff. after PCA)	Feed forward input delay NN *92* %
14.	D. Lederman [**80**]	1. Normal *vs.* RDS 2. Normal *vs.* cleft palate	*21* healthy *19* RDS *7* cleft palate Total *1723* records	MFCC	1. *63* % *2. 90* % With CD HMM
15.	Alalie [**4**]	Normal *vs.* pathology Bovine protein allergy Tetralogy of fallot Thrombosis in the vena cava Cardio complex X chromosomal abnormalities Coarctation of aorta	Normal (*38+25*) Pathology (*13+5+13+9+14+9+10*)	MFCC	*87.3* % Adapted boosted mixture learning method for GMM
16.	Sergio D. Cano Ortiz [**81**]	Normal *vs.* pathology	*19* healthy *16* sick Sick: hypoxemia retarded intra-uterine growth hypoxemia with another risk hyperbilirrubin	*25* features prosodic	ANN 85 %

2.9 Literature Search on Sudden Infant Death Syndrome (SIDS)

Research work for the study and understanding of sudden infant death syndrome (SIDS) phenomenon is a challenging task. SIDS is caused by maldevelopment or delay in maturation of the brainstem neural network that is responsible for arousal and affects physiological responses to life threatening challenges during sleep. Prone sleep position in infants is also considered as a possible cause of SIDS. Another study shows that in this position decreased arousal and response to pain are responsible factors for SIDS [**82**]. Infant groups which are at high risk of SIDS are low birth weight, premature infants, infants experiencing an idiopathic apparent life threatening event and siblings of SIDS. Prolonged sleep apnea is a contributing factor in SIDS. Colton and Steinschneider found that cries of SIDS infants are longer with a low F_0 and lower formant frequencies compared to healthy term (HT) group [**83**]. Corwin showed that there is no difference in cry duration in the two groups. In addition, the cries were found to be of higher F_0 and higher formant frequencies [**84**]. In the study of SIDS, siblings of SIDS infants are considered. In one of the work reported in [**85**], SIDS infants sibling's cries are compared with the normal infant cries and the cries are analyzed using the features such as overall cry duration, cry ratio (expiratory cry to total cry duration), cry latency, first spectral peak, spectral tilt (ratio of energy (*0-1000* Hz) to energy (*1000-5000* Hz), high frequency energy (energy in *5-8* kHz). It was shown that the spectral tilt and first spectral peak differs in SIDS sibling group (both values are higher than healthy term (HT) group). SIDS group show higher latency than HT group. Siblings group showed higher F_0. The factors of high risk include young maternal age, birth rank, poor social conditions, low birth weight, asphyxia and neurological damage at birth, subsequent siblings of SIDS victim and cardio-respiratory abnormalities such as prolonged QT internal, excessive periodic breathing, and prolonged sleep apnea [**86**], [**87**]. In a recent study, it is found that the cries of SIDS infants are longer and have a high difference in the absolute values of first and second formants [**88**].

Based on the literature presented in this chapter and summary of the contributions (as described in Section 1.6), this thesis presents work on the following aspects:

1. Infant cry data collection and corpus design
2. Investigating various signal processing challenges associated with applying traditional and state-of-the-art signal processing algorithms directly to infant cry signal.

3. Detailed spectrographic analysis (including exploiting ten different cry modes, selection of window length and window type in STFT computation, justification of the use of narrowband spectrogram opposed to the use of wideband spectrogram.
4. Analysis of infants at high risk (who are suffering from respiratory distress and hence, may be prone to SIDS).
5. Use of higher-order spectra such as bispectrum with higher-order singular value decomposition (HOSVD) for feature dimension reduction for infant cry classification.
6. Classification of asthma and HIE pathological infant cries.

2.10 Chapter Summary

In this chapter, physiology of cry production and the role of CNS in the cry production and control is explained. Dependencies of various factors which affect the cry behaviour in infants are discussed. Along with this, cry analysis methods such as time-domain, frequency-domain, spectrographic analysis, computer-based algorithms are explained along with their comparative advantages and disadvantages. The recent work in the proposed area is also detailed with the limitation that their work cannot be compared because all the researchers are working in this area are using their own datasets and each of these studies are dealing with different set of pathologies and cry types. Several possible fields of research in the infant cry processing are explored with the latest work in that direction and the limitations of the present work.

In the next chapter, details of the corpus collected for infant cry analysis are given. Procedure for data collection and ideal characteristics of the dataset are given. Other databases which are also used in this thesis work are also detailed in the next chapter.

Chapter 3.

Data Collection and Corpus Design

3.1 Introduction

Analysis of infant cries may help in identifying the needs of the infants such as hunger, pain, sickness, *etc.* and thereby developing a possible tool or mobile application which can help the parents in monitoring the needs of their infant. Analysis of cries of infants who are suffering from the neurological disorders and severe diseases (which may, later on, result in motor and mental handicap) may prove helpful in early diagnosis of these pathologies and possibly preventing infants from such disorders. For the analysis of infant cries and for the development of such tools, development of an infant cry corpus is necessary. Infant cry database is not available commercially for research which limits the scope of research in this area. Because the cry characteristics changes with several factors such as reason of crying, infant's health and weight, age, *etc.*, care is required while designing a corpus for a particular research application of infant cry analysis and classification.

In this chapter, ideal characteristics of the corpus, factors influencing cry characteristics are proposed and author's experiences during data collection are shared. This study may help other researchers to build infant cry corpus for their specific problem of study. Justification of the proposed characteristics is also given along with suitable examples. In addition, description of corpora used in this thesis is also given with their statistics and metadata information.

3.2 Need for Infant Cry Database or Corpus

Cry is the first response of an infant as a result of interaction with the external world. Till an infant learns to speak, it is the only mechanism to communicate with which he or she can express his or her needs and ask for attention. Infant cry is an important signal for analysis of an infant's health condition. Analysis of infant cries may help in identifying the needs of the infants such as hunger, pain, sickness, *etc.* and thereby developing a tool or possible mobile-based application which may help the parents in monitoring the needs of their infant. Analysis of cries of infants who are suffering from the neurological disorders and severe diseases which can, later on, result in motor and mental handicap, may prove to be helpful in early diagnosis of these pathologies and prevent infants from such disorders.

The main challenge in the infant cry analysis and classification is the unavailability of the statistically meaningful database (or at least sufficiently large number of infant cry samples). Collecting a database requires permissions from the hospital authorities and parents as well. Getting cry signals of infants suffering from pathology is furthermore difficult. Getting a statistically significant corpus is a challenging task. Most of the researchers working in this area have their own database with different sets of infant cry types, recording conditions, microphones, age groups, different pathologies and different weights of infants. Standard database for the task is not available which also restricts the comparison of different research works.

Cry signal characteristics changes with several factors such as reason of crying, the age of the infant, *etc.* in this area. In pathological cry analysis, cry characteristics changes with the severity of the disease. In such cases, long-term follow up of the infant is required which is a time-consuming and difficult task. All these effects altogether pose a challenge to the researchers to work in this area and contribute towards it. In this chapter, the effect of these factors on the infant cry analysis is presented and the general guidelines for data collection and corpus design are suggested so that the researchers who are interested to work in this area can collect their data accordingly and design their corpus which can be used potentially in their research work. This study will help in possible standardization of infant cry corpus preparation.

3.3 Ethical Issues in Infant Cry Data Collection and Protection of Human Rights

Infant cry recording is a very sensitive task involving dealing with newborn human participants. Researchers have considered pain cry and hunger cry for infant cry analysis. To record the pain cry, how stimulation should be given to the infants is a debating issue for a long time. In the medical-domain, pain cry is recorded by giving a rubber snap on infant's foot. As far as this is done by a medical practitioner, parents generally do not object. However, for others, they do not allow to do it on their child. Even hospital authorities also do not allow using such practice for data collection purpose. Data collection of pain cry is possible only while *immunization process* or during treatment if the injection is given. Data collection of pain cries during immunization process give several advantages such as the amount of stimulation is controlled (which is not the case in other methods), cries collected will be of the similar age groups as for a particular vaccine, specific age is defined and generally, in immunization units, silence is maintained. Recording of hunger cry is even more difficult. For data collection of hunger cry, the only option is to wait until infant gets hungry. Another important issue is to convince the parents for the purpose of data collection and getting their approval. For normal infants, this part is comparatively easy. However, when an infant is sick, especially, those who have severe pathologies, this task become very difficult. In these cases, emotional status of the parents should be taken care and all their decisions even if it is not positive for data collection should be respected without hurting their sentiments.

To look into these sensitive issues such as the method of cry stimulation and protection of human rights, Institutional Ethics Committee

(IEC) is formed by hospital authorities. The objective of the IEC is to ensure a competent and consistent ethical review mechanism for health and biomedical research proposals dealt by the committee as prescribed by the ethical guidelines for biomedical research on human participants (Indian Council of Medical Research (ICMR) [**89**] in India or Council for International Organizations of Medical Sciences (CIOMS) and World Health Organization (WHO) guidelines [**90**] or respective country's guidelines). The ethical committee ensures that the procedures used for scientific research on human participants do not harm the participants under study. It looks into the feasibility of the research area and the methods used in the study for research and its application for future use. The composition of IEC is multidisciplinary and it includes experts from several disciplines such as medical, legal, social welfare area, lay person and clinician from different institutes to have independence in the composition of the committee.

The research proposal for infant cry research and data collection should be reviewed by the IEC and the consent for data collection needs to be taken before collection of infant cries. During data collection, general principles in biomedical research involving human participants should be followed as mentioned in "Ethical Guidelines for Biomedical Research on Human Participants" described by Indian Council of Medical Research, New Delhi or respective hospital's guidelines (for other countries). Based on these recommendations, few important guidelines for the infant cry data collection are as follows:

a. The participation of infants must be voluntary in nature. No one can be forced to be a part of this study.
b. Before participation in the data collection, parents must be informed about the purpose of the study and the method of data collection. Written consent must be obtained from the parents (because the participants are minor in this study).

c. The data and associated metadata of the parents and infants who have participated in the study should be kept confidential.
d. Data can be collected using a hand-held recording instrument to ensure minimum risk or non-intrusion to the infant.
e. Data should be collected by the researcher to avoid chances of mishandling of the data and privacy of the participants.
f. Data should be collected in the presence of medical practitioners to avoid any possible harm due to accidents and get their feedback on the research work for normal *vs.* pathological cases.

3.4 Metadata Preparation

Before collection of the infant cry data, parent consent form and participant information form must be prepared with due care. The parents consent form must indicate the purpose of the study, method of data collection, terms of compensation in case of any injury occurred during data collection and privacy conditions. For the preparation of the subject (infant) information form, the purpose of the study is very important. For example, in normal healthy infants, the cry acoustics are dependent on the reason of crying. However, cry acoustics are dependent on the pathology and its severity in pathological cry analysis. Since, it is difficult to collect data from infants, it is advisable to get as much information about the infant in the participant information form as possible, so that the database can be used for several purposes. The details required in the participant information form are: name of infant, date of birth, weight at the time of birth, gestation age (GA) at the time of birth, current weight, reason of crying, gender, details of siblings, history of any sibling deaths, parents name, any genetic disease to any of the parent, their educational qualification (in some studies of child development it has been observed that infant's language development is directly related to his mother's educational qualification) and date of recording with *.*wave* file name. In the case of sick infants, details of disease and doctors comments about the severity of disease are necessary. A template for metadata preparation (Subject Information form) is shown in Appendix A.

3.5 Ideal Characteristics of Infant Cry Corpus

The desirable characteristics of the infant cry database are as follows:

a. The reason of crying should be similar in a particular class for the infant cry classification task.
b. The age group under consideration should be similar.
c. The weights of the infant should not vary too much.
d. In the study of premature infants, gestation age (GA) is important to consider for infant cry analysis.
e. During data collection of pain cries, pain stimulation must be similar for all infants.
f. The number of cries, as well as the number of infants, should be significantly high to give statistical meaning to the findings or results.
g. The cry utterance should be long for the study of reasons of crying.
h. Sampling frequency (F_s) should be kept high.
i. In neonatal infants (*0-1* month), cry characteristics changes very frequently due to the rapid development of the respiratory system and its coordination with central nervous system (CNS). This group can be considered separately.
j. Infants having less than *3* months of age are obligate nose breathers, this group of infants can be dealt in a different class from their older infants. After the age of three months, infants have separate paths for food and air (for breathing) and hence, it becomes more like an adult system. Thus, they form a separate group in the cry analysis.
k. In pathological cry analysis, number of participants should be high to give a better understanding of the reason of crying. Collecting more cry

samples from very few infants (participants) may not give statistically significant / reliable results.

l. In the study of pathological cries, the reason of crying should be the same. The reason of crying will give changes in acoustic features in addition to the changes in acoustic features due to presence of pathology.
m. Infants, who have a history of sibling's deaths, their cries can be analyzed separately. Furthermore, such infants are considered as high risk infants in the study of SIDS.

3.6 Efforts and Experiences during Data Collection

Some efforts and experiences during data collection are as follows:

a. On the day of recording, concerned doctors and concerned staff were also informed about the purpose of the study and they had been very cooperative during the recording of cry samples and in describing the pathology of the infants. They tried to maintain silence as far as possible.
b. Sometimes, it was difficult to convince the parents to give consent for recording. However, no one was forced to give permission for cry recording rather recording was done for only those infants whose parents voluntarily agreed for this task.
c. Because of the presence of *4-6* doctors and several infants and their mothers present in outdoor patient department (OPD) room, sometimes noise-level was very high. Such cases were ignored in the analysis.
d. At times during the recording of an infant's cry, the cry of another infant present in the same OPD was dominating. It made the recording task very difficult. Such cases were excluded while recording. In the case of too much noise, it's better to stop recording the infant cries.

Instrument settings can be adjusted and it should be placed near to the infant's mouth in order to minimize the surrounding noise effects to a certain extent.

e. Sometimes as soon as voice recorder was taken near to the infant for the recording of cry, infant keeps quiet and was found to start observing the recording instrument. Such cries are also discarded from the analysis as a transition in infant's behaviour changes the prosodic content of the cry.
f. Almost all the infant cries were spontaneous. For getting a cry sample, *none* of the infants was given any external stimulation for data collection purpose.

3.7 Database Collected and Used in the Research

In this work, to show the results and illustrations, we have used two databases which were collected from various hospitals of India. While the recording of the cry, doctor's comments were also recorded. One more database is borrowed from other researchers called Baby Chillanto database. Database collected by the researcher is referred as *Corpus I* in this work [**91**]. Data was collected from the Civil Hospital, Ahmedabad, India, the biggest hospital in Asia and one of oldest and modern hospital in India [**92**]. For data collection, permission has been granted by the hospital authorities. Data were collected under the observation or supervision of senior doctors and resident doctors. Data was collected from the pediatric outdoor patient department (OPD) and pediatric ward. For data collection, Zoom *H4n* portable recorder was used. It has a *24*-bit quantization and *96* kHz adjustable sampling frequency. In our experiment, we kept sampling frequency at *44.1 kHz*. All signals were recorded in stereo mode in **.wav* format. Data is then transferred through a USB cable to a laptop. Data was collected from OPD unit where there are at a time *4-6* doctors observing the patients. The data was recorded in real-life noisy environments. As far as possible, noise effect has been tried to suppress by reducing the area coverage range of recording instrument. The recorder was kept at a distance of *5-6* cm away from the infant's mouth to further reduce noise effect. In all the cases, the infant was sitting in his/her mother's lap.

The other database collected by Ms. Neeharika Buddha is referred as *Corpus II* [**93**]. Infant cry data in this database was collected from three places, namely, 1. King George Hospital, Visakhapatnam, India (from this hospital, data of neonates was collected), 2. Prabha Nursing Home, Visakhapatnam, India (from this hospital, first cries of newborns were recorded) and 3. Child Clinic, Visakhapatnam, India (from the child clinic cries of sick infants and pain cries of normal infants during vaccination were recorded). *Corpus II* was collected with a portable Cenix digital recorder with an external microphone. The sampling frequency of the recording was *12* kHz and it was quantized at *16*-bits PCM. Most of the recordings were done in a silent area. However, in some cases (especially, in the case of sick infants) sounds of parents pampering the infants were also recorded (which was naturally unavoidable). The corpus statistics are shown in Table 3.1-Table 3.10.

Table 3.1. Distribution of infants

S. No.	Type	*Corpus I*	*Corpus II*
1.	Normal infants	93	61
2.	Pathological infants	56	26

Table 3.2. Distribution of samples over gender

S. No.	Gender	*Corpus I*	*Corpus II*
1.	Boy	114	38
2.	Girl	60	49
sometimes more than one cry samples are collected from a single infant.			

Table 3.3. Detailed distribution of samples over gender for *Corpus I*

	Normal	Pathological	Pre-term	Fullterm	Pain	Hunger
Boy	30	20	6	44	27	23
Girl	31	6	3	34	25	12
Total	61	26	9	78	52	35

Table 3.4. Distribution of samples over cry type for *Corpus I*

	Hunger	Pain
Normal	17	45
Pathological	32	8

Table 3.5. Distribution of samples over age for *Corpus I*

Age	Number of samples
1 day	23
< 1 week	38
>1 week and < 1 month	31
1-6 months	37
6 months- 1 year	22
1 year – 1.5 years	13

Table 3.6. Distribution of samples over age for *Corpus II*

Age	Total
1-7 days	6
<1 month	7
1-4 months	37
4-8 months	20
8 months-1 years	14
1 year – 2 years	3

Table 3.7. Distribution of samples over pathologies for *Corpus I*

Pathology	Total infants
Brain hemorrhage	1
Cleft lip	2
Diarrhea	2
Down syndrome	2
Gastritis	3
Heart disease	2
Hydrocephalus	2
Hypo calcium	1
Mal nutrition	2
Pneumonia	2
Pyomeningitis	4
Upper respiratory tract infection	8
Bronchitis	2
Jaundice	1
Epilepsy	2
Thelesimia	3

Table 3.8. Distribution of samples in special cases for *Corpus I*

Case	Total infants
Siblings Death History	9
C section Delivery	30
Normal Delivery	57

Table 3.9. Distribution of cry recordings in different health conditions for *Corpus II*

Class	Number of cries
Newborn Normal birth	45
Newborn preterm	36
Normal healthy cry	93
Pathological	56

Table 3.10. Distribution of pathological cry samples among several pathologies in *Corpus II*

Pathology	Number of cries
Asthma	7
HIE	14
Hyper bilurubin	4
Meningitis	4
Respiratory distress	10
Miscellaneous cry (fits, heart disease, broken bones, larynx not developed, jaundice, cleft lip, high risk baby)	12

Another database used in our work for some experiments in few sections of this thesis, is the Baby Chillanto database. In this thesis, this corpus is referred as *Corpus III.* The Baby Chillanto Database is a property of the Instituto Nacional de Astrofisica Optica y Electronica-CONACYT, Mexico. We like to thank Dr. Carlos A. Reyes-Garcia, Dr. Emilio Arch-Tirado and his INR-Mexico group, and Dr. Edgar M. Garcia-Tamayo for their efforts and dedication towards the collection of the infant cry database. This database contains two kinds of samples, one contains complete recordings of individual crying, and the other has the same samples divided into one second duration segments. The cry samples were collected by medical doctors from several specializations in pediatric or specialized hospitals. The normal infant cries have hunger and pain recordings. The other samples are the cry recordings of the deaf infants and infants suffering from asphyxia [**54**]. This database is used for comparison of normal and other pathological signals of infant cry from the cries of infants who are deaf or suffering from asphyxia. The statistics of the database is given in Table 3.11 and Table 3.12.

Table 3.11. Number of cry recordings of normal, deaf and asphyxia cries in *Corpus III*

S. No.	Health	Number of Recordings
1.	Normal	21
2.	Deaf	52
3.	Asphyxia	6

Table 3.12. Number of cry samples (*1s* duration) in various health conditions in *Corpus III*

S. No.	Health	Number of Samples
1.	Normal	507
2.	Deaf	879
3.	Asphyxia	340

3.8 Factors Influencing Infant Cry Characteristics

The ideal and desirable characteristics of the infant cry corpus are mentioned in Section 3.5. In this Section, justification of these characteristics is given with suitable examples using earlier studies on infant cry analysis and narrowband spectrograms which represent the joint time-frequency energy density of infant cry signal. Spectrographic analysis has been used in infant cry analysis by the several researchers, *e.g.*, [**39**] and [**94**]). Several factors which influence the cry characteristics need to be considered while collecting the infant cry signal for corpus preparation. Some of these factors are as follows:

3.8.1 Variability of Acoustic Features with Age (newborn to *1* year)

An important factor to be considered in the analysis of infant cry is variability of acoustic features with age. The effect of acoustic features is shown in Figure 3.1- Figure 3.3 using cry samples taken from *Corpus II* cry samples. In these figures, the spectrograms of the infant cry signals of different ages, *i.e.*, *2* months, *6* months and *12* months old are shown. All the cry signals are of normal infants for the same reason of crying (pain). From the spectrograms, it is observed that as the infant grows older, the cry duration increases. In younger infants, the pauses between the cries are longer and cries are of short duration. As the infant grows, he or she learns to control the respiratory movements and have better neural control. Because of which, the cry duration becomes longer. In neonates, the respiratory rate is *40 rpm* (respirations per minute), which reduces to *30 rpm* in infants of age *12* months whereas respiration rate is *20 rpm* in the case of adults. Another important observation is that the power in the cry signal increases with the growing age. It also shows the improvement in the muscular strength of the voice production system. It has already been reported by the researchers that with the growing age of the infant, pitch and formants of the cry signal decreases due to increase in vocal tract length.

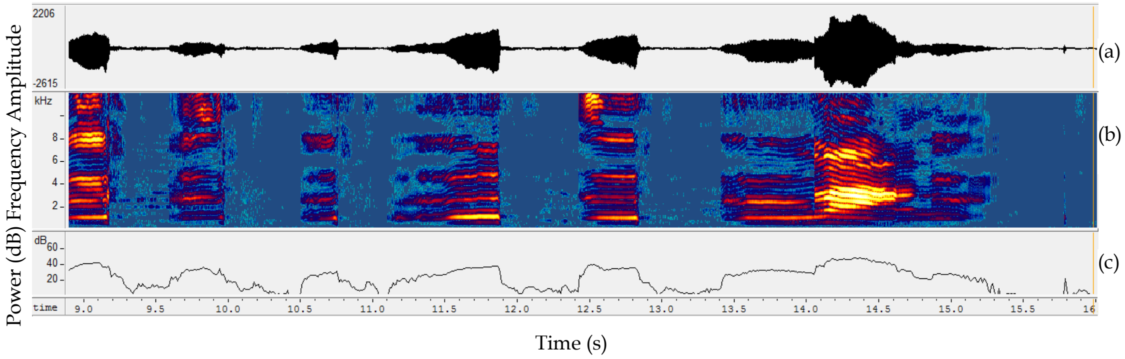


Figure 3.1. Spectrogram of a *2* months old infant's cry sample (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a).

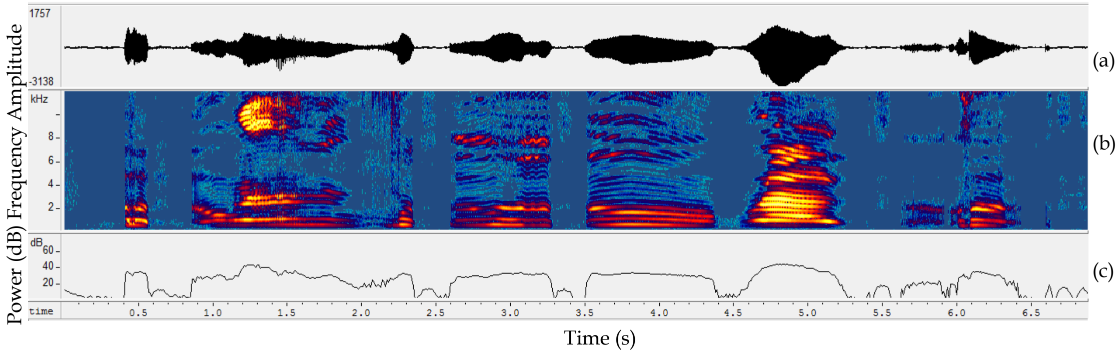


Figure 3.2. Spectrogram of a *6* months old infant's cry (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a).

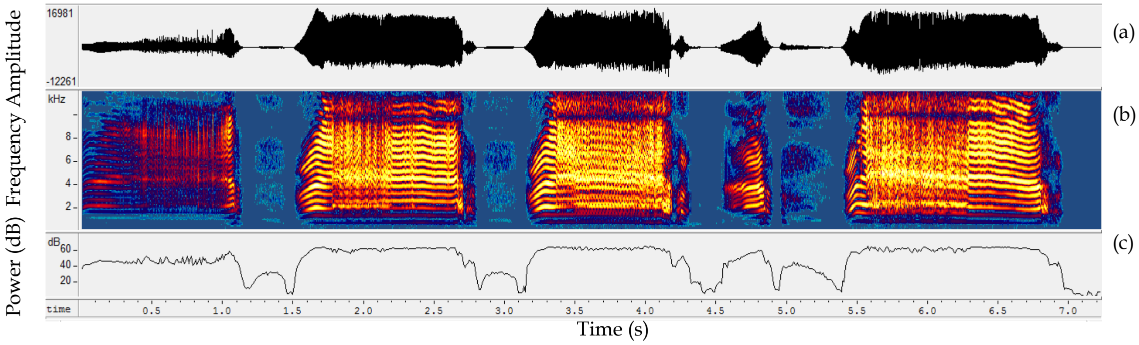


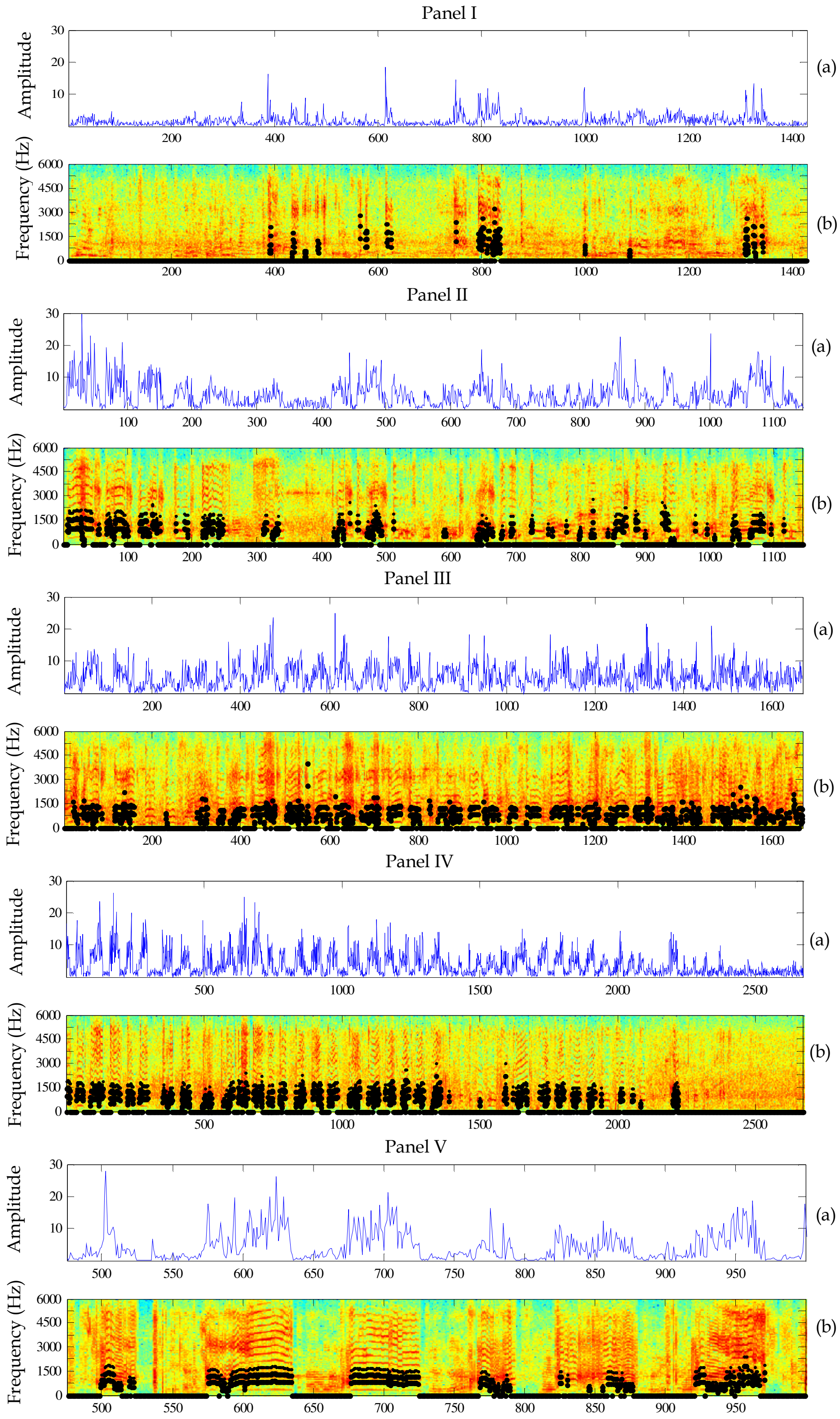
Figure 3.3. Spectrogram of a *12* months old infant's cry sample (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a).

All these changes in the cry pattern cause difficulties in the analysis of the infant cry. However, for adults, the change in the fundamental frequency (F_0), formants and duration features for the same pronunciation are comparatively small (because of well-defined rules for the pronunciation of a particular phoneme). Effect of the development of the speech production system has been observed in our analysis of an infant whose cries were recorded for *20* days from the day of birth (from the same infant). The results of analysis are shown in Table 3.13, where the mean fundamental frequency (mean F_0), maximum fundamental frequency (max. F_0) and unvoicing ratio (VUV ratio, which is the ratio of number of unvoiced frames to total number of frames in a cry) of the cry is calculated from the cry recording. The ratio of unvoicing in the cry is also calculated to show the development of the vocal olds.

Table 3.13. Variation of acoustic features with age on *Corpus I*

Age in days	Mean F_0	Max F_0	VUV ratio
0	428.57	800.00	0.95
1	393.44	857.14	0.67
3	400.00	1333.00	0.59
10	406.60	750.00	0.71
15	406.78	1090.9	0.64
20	400.00	857.14	0.62

From Table 3.13, it can be observed that the mean F_0 of the infant cry reduces with the development of the vocal system. However, with the maturity of neural control system and speech production system, infant learns to modulate the cry and can increase the fundamental frequency (F_0) or pitch upto *1* kHz. These high-pitched cries are used by the infants to draw the attention of the caretaker in case of emergency. The voicing content in the cry also increases with the age and cries become rhythmic. From Figure 3.4, it is observed that with age the respiratory control of infant becomes better and it results in longer duration of cries. This also results in higher energy in the cries with increasing age. Along with this, instead of having abrupt changes in the energy at the inspiratory durations, the energy transitions become smoother (compare newborn and *20th* day cry signal energies for the same infant). Moreover, we can observe vibration pattern in F_0 contour as compared to only rising pattern which is visible in the first day cry.


Panel I

Panel II

Panel III

Panel IV

Panel V

Figure 3.4. Variation of acoustic features in infant cry with age (F_s= *12* kHz). Panel I: birth cry, Panel II: Day 1 cry, Panel III: day 4 cry, Panel IV: day 11 cry and Panel V: day 20 cry.

In all the subfigures: (a) represents short-time energy of the cry signal, where X- axis is the frame index and Y-axis is the signal energy, and (b) shows the spectrogram and fundamental frequency (F_0) contour of the cry superimposed on the spectrogram of the cry where X- axis is the frame index and Y-axis is the frequency in Hz. All samples are taken from *Corpus I.*

3.8.2 Variability of Acoustic Features with Weight of the Infants

Cry characteristics in infants with different weights follow the same trends as that of age. Infants with higher weight in the respective age group show dominant prosodic marks in their cry. From the *10*th day after birth, when the post-birth weight loss is usually regained, there is a steady increase in weight so that during the first three months an average baby gains about two pounds per month, or nearly one ounce per day. At five months, the birth weight is doubled. Beginning at six months, there is only a one pound gain per month in weight so that the birth weight is tripled at the end of the first year and quadrupled at the end of the second year [**95**].

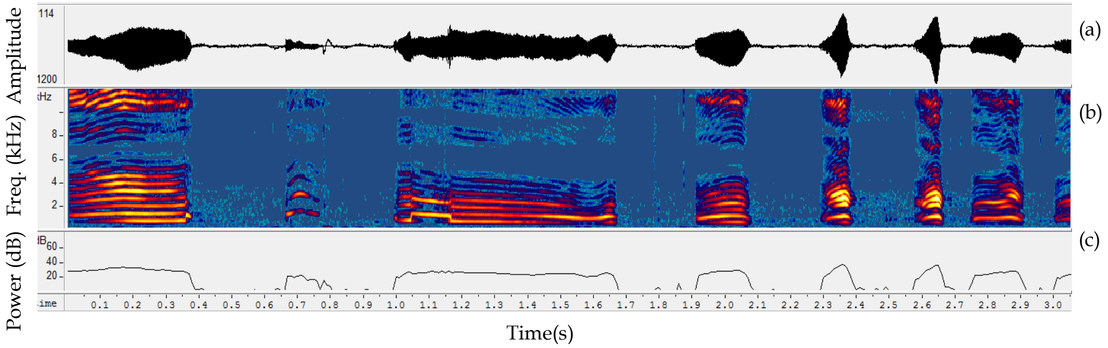


Figure 3.5. Infant of *2.1* kg weight of age *4.5* months of age (F_s= *44.1* kHz). (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a).

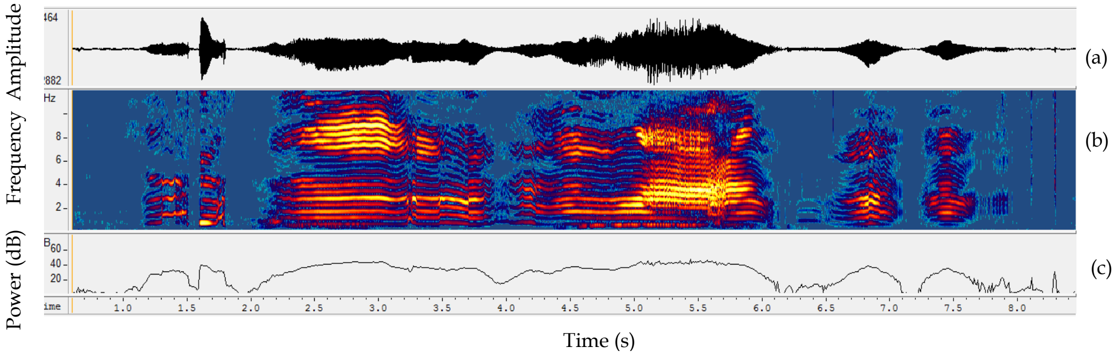


Figure 3.6. Infant of *6.2* kg weight of age *5* months of age (F_s= *44.1* kHz). (a) time-domain signal, (b) narrowband spectrogram and (c) power plot of (a).

Figure 3.5 and Figure 3.6 show the variability in the acoustic features for the infant cries of two infants with almost same age, however, with different weights (samples from *Corpus II* are used). Comparing the spectrograms of the low weight and high weight infants (especially, in the present case, the weight difference is very high), it can be observed that for low birth weight infant, the duration of cry is small and it has poor control on the respiratory system. F_0 is also high in the low weight infant. In premature low birth weight infants, the pitch is found to be higher than the normal full-term infants. The power content and prosodic variations are high in high birth weight infant. Improvement in the weight of the infant is correlated with the development of neural, muscular, and anatomical structures. It is also an indication of the integrity of the various anatomical structures with the neural system. In the case of adults, because the voice production system is fully developed, the variations in weight of the participants do not change the results of the speech analysis or recognition. However, recently, an attempt has been made to estimate the height of a person from the state-of-the-art spectral features, namely, Mel frequency cepstral coefficients (MFCC) [**96**].

3.8.3 Cry Type Differences

The cry characteristics also change with the reason of crying. The researchers working on the infant cry analysis divided the cry types into four types, namely, birth, hunger, pain and pleasure. The characteristics of the cry such as durational features and fundamental frequency (F_0) based-features vary with the type of cry. For example, the mean maximum and minimum pitch for the hunger cries are *550* Hz and *390* Hz, whereas these parameters are *650* Hz and *360* Hz, respectively, for pleasure cries. On the other hand, in birth cries these are *550* Hz and *450* Hz, for pain cries the values of maximum and minimum pitch are *650* Hz and *400* Hz, respectively [**11**]. Sometimes, infants also cry because of fatigue, however, these cries are very difficult to identify and collect. Hence, these are ignored in our analysis and most of the literature which deals with signal processing aspect of infant cry does not have an analysis of this cry type.

3.8.4 Anatomical Differences in Airways

Infants have a proportionately larger head and tongue, narrow nasal passages, an anterior and cephalad larynx (at a vertebral level of *C3-C4*), a long epiglottis and a short trachea and the neck. Vocal folds in infants are *3-5* mm long and the composition of the vocal folds is uniform. The vocal folds in infants are much smaller than vocal folds of adults and they lack lamination seen in the adults. This lamination plays an important role in the theories of phonation studies. Vocal fold length reaches to *7.5 mm* by the age of *5* [**97**]. These anatomic features make neonates and most young infants obligate nasal breathers until about *3-4* months of age. The cricoids cartilage (subglottis) is the narrowest point of the airway in children. All these anatomic differences make signal processing of the infant cry difficult than the adult speech signal. The position of the larynx in the infants is close to the base of the skull. This high position of larynx helps in forming a closed passage from nose to the lungs. The newborn infants can move the larynx upward in the nasophagus. The soft palate and the epiglottis affect a double seal and liquids can flow through the larynx while air flows through the nose through the larynx and through the trachea down to the lungs. There is *no* possibility of choking by having lodge into the larynx as is the case with the adults. Moreover, infants are obligate nose breathers, *i.e.*, they do not breathe through the mouth even in the case of nose blocking [**24**].

Shorter length of vocal tract results in high formant frequencies as they have an *inverse* relationship (as shown in eq. (4.3) in Section 4.2) [**98**]. Normally, the first two or three formants are studied in infant cry analysis. The first two formants of infant cry are observed at *1100* Hz and *3000* Hz, respectively. This can also be approximated from the vocal tract length of the infants which is approximately *7* cm in length. The tongue in the newborn is long and thin compared to adult human being. The tongue is positioned in the oral cavity and does not have almost circular shape. This difference in the tongue shape makes it impossible for a newborn to produce supralaryngeal vocal tract area function that is necessary to produce sounds [**11**].

3.8.5 Variability of Acoustic Features with Gestation Age (GA) in Premature Infants

In premature infants, the cry characteristics changes with the gestation age (GA). GA is the age of the infant counted from the date of conception of the fetus and is measured in weeks. A normal pregnancy range from *38-42* weeks and an infant born before *37* weeks are considered premature. It has been shown that infants born with *31-33* weeks of GA have a smaller duration of cries (~*1.2* sec) compared to infants born with *38-41* weeks of age (~*2.6 sec*). The mean maximum pitch and mean minimum pitch values are also higher in infants with low GA. In infants with *31-33* GA, the mean minimum and maximum F_0 values are *990* Hz and *470* Hz, respectively. The values of these parameters are *750* Hz and *370* Hz, respectively, in infants born with *38-41* weeks of GA [**99**].

3.8.6 Variability of Acoustic Features with Pathologies

All pathologies effect differently to different organs in the infants as well as in adults. In the spectrographic analysis of the cries, various differences in the cry modes are observed for different pathologies. Some of the spectrographic analysis of the pathological infant cries is given below.

3.8.6.1 Cry of an infant suffering from laryngomalacia

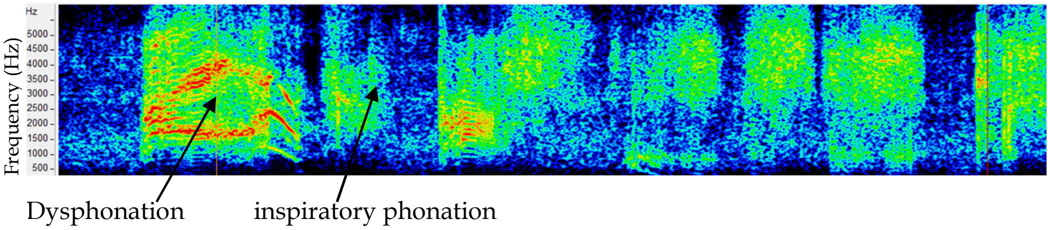


Figure 3.7. Spectrogram of an infant cry suffering from laryngomalacia taken from *Corpus II.*

Spectrographic analysis: It is observed from Figure 3.7, that

1. Dysphonation and inspiratory phonation are dominating.
2. Double harmonic break, glottal roll, glide are totally absent.
3. Spectral resolution is poor.

3.8.6.2 *Cry of an infant suffering from asthma*

Spectrographic Analysis: Due to frequent inhalation, inspiratory phonation is observed in the spectrogram as shown in Figure 3.8. The double harmonic break is visible in the spectrogram of the cry.

1. Because of the problem in breathing, inspiratory phonation is frequent in the spectrogram.
2. Rising and falling modes are present. It is similar to normal infant cry.

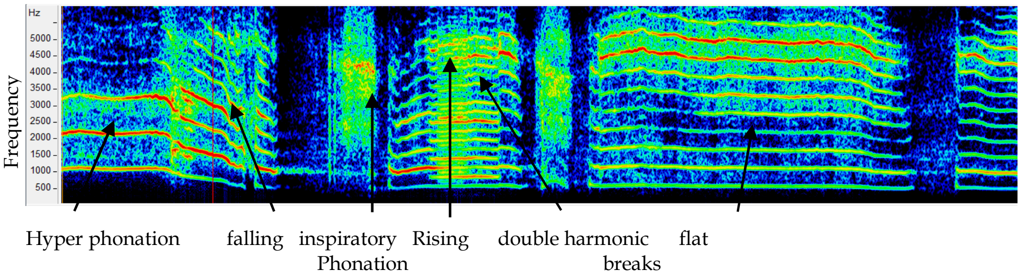


Figure 3.8. Cry modes present in the spectrogram of an infant cry suffering from asthma sample taken from *Corpus II.*

3.8.6.3 *Cry of an infant suffering from congenital heart disease*

Spectrographic Analysis: Spectrogram of an infant cry suffering from congenital heart disease is shown in Figure 3.9. It can be observed that the melody type is rising followed by falling, same as a normal infant. Glottal roll is present in the spectrogram and dysphonation is dominating in the spectrogram.

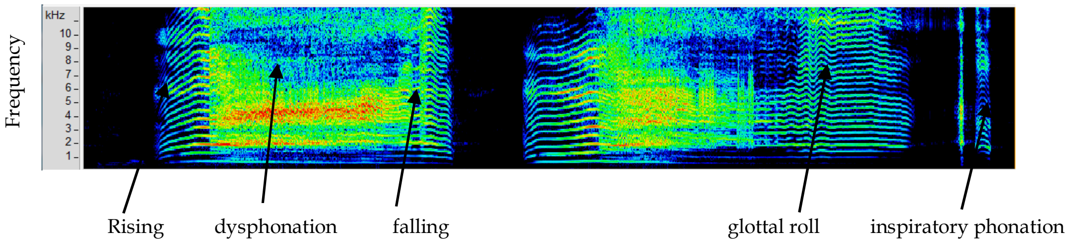


Figure 3.9. Cry modes present in the spectrogram of an infant suffering from congenital heart disease sample taken from *Corpus II.*

3.8.6.4 *Cry of a normal infant*

Spectrographic Analysis: Spectrographic analysis of the normal infant cry shows that the infant cry has rising cry melody followed by falling melody pattern. Inspiratory phonations are visible in the cry. However, the duration of the inspiratory phonations is smaller than the pathological cases.

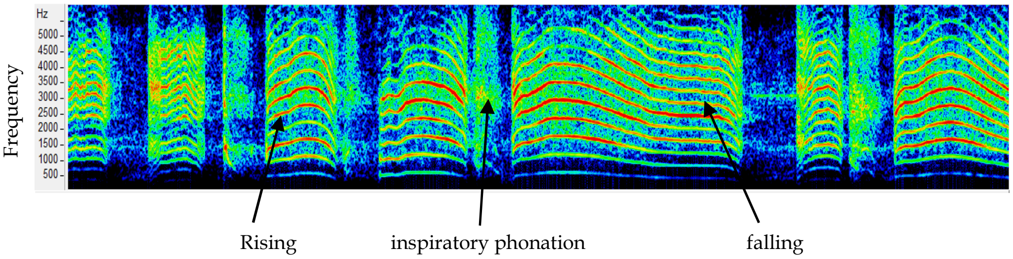


Figure 3.10. Spectrogram and cry modes present in the spectrogram of a healthy normal infant's cry sample taken from *Corpus I*.

From the above analysis of the spectrograms of normal and pathological infant cries, it can be observed that pathological cries have vibrations, double harmonic breaks (many times also correlated to the muscular pain [**40**]) and dysphonation (*i.e.*, unstructured energy distribution) modes in their spectrograms. A Higher percentage of dysphonation, the presence of double harmonic breaks and glottal roll may be attributed to the presence of some disorder, though these are not yet found to be specific to any disease or pathology. Thus, which pathologies should be considered for the analysis of the infant cries during data collection is an important aspect. In specific cases such as deafness and asphyxia, differences in the cry patterns are observed compared to normal infants, even through auditory inspection of the infant cries. However, it does not give reliable classification or distinction among these cries. Pathologies which are not due to neurological disorder, the class-specific features in the cry may not reflect in cry acoustics. Thus, a detailed study of the pathologies which are to be considered in the analysis is compulsory in order to avoid wrong selection of pathologies. On the other hand, pathologies which reflect differences in the infant cry acoustics are difficult to find among infants. Thus, creating a statistically significant database is a challenging task. In such cases, suggestions from medical practitioners and nurses can be proved tremendously helpful before starting data collection. The detailed spectrographic analysis of normal and pathological infants is presented in Chapter 5.

3.8.7 Birth Cry or First Cry Analysis

Birth cry is considered as a separate cry type because it conveys important information about the health of the infant. The birth cry is a symbol of the beginning of a healthy life. This is the first response of the newborn to the external world after coming out from the liquid atmosphere of the uterus to air. It ensures proper functioning of the lungs and this is the first time lungs full up with air and expands to its fullest capacity. It also helps the babies to get rid of any amniotic sac present in the lungs and nasal cavity [3]. After the birth, if an infant doesn't cry, it implies that something may be wrong with an infant and that infant has to be rigorously investigated by the pediatrician. Pediatricians use the apgar score to test the health of the newborn. To evaluate the apgar score, five parameters (namely, complexion, pulse rate, reflex, activity, and respiration) are rated on a *0-2* scale and they are summed to get the apgar count. Infants with score *7* or above are normal healthy infants and those with score *3* or below are regarded as critical. The detailed study of newborn's cry analysis is presented in Chapter 5.

3.8.8 Stimulation for Cry Production

As discussed in Section 3.8.3, during collection of the pain cries of infants, the important question is the how the stimulation should be given to the infants to elicit pain which can result in crying. The procedure should be such that it follows the human ethics, guidelines and acceptable to the parents of the infants as well. The amount of stimulation should be maintained same for all the infants during data collection. Best possible solution to this is to collect infant cry data during vaccination. It has two advantages, namely, pain stimulation is not objectionable by the parents and it is fully controlled by the experienced practitioner. Secondly, infants of same age group will be covered in the analysis which adds accuracy and consistency to the results of the research.

3.8.9 Room Acoustics

The environment of data collection should be as silent as possible. However, it is difficult in infant cry analysis because in hospitals many infants may cry at the same time. Moreover, parents may interfere in the recording in order to sooth their babies. Hence, efforts must be made to minimize the noise.

3.8.10 Time and Duration of Infant Cries

In the case of pain cries, the recording should be done as soon as the stimulation is given to elicit the infant cry. In the case of pathological cries (neurological disorders), it has been noted that the latency (duration between stimulation and cry production) is higher compared to normal health infants. This feature may help in classifying the pathological infant cries. It is always better to have a complete recording of a cry utterance (from start to end when the infant stops crying), in order to study the cry behaviour using prosodic features. This analysis may help in identifying the reasons of crying.

3.8.11 Recording Device

A small hand-held voice recorder (with the high sampling frequency and excellent bit depth or resolution) is preferred so that one can move it closer to crying babies as soon as he or she cries. Recent advances in recording instruments have made available, very small size microphones which can be attached to the infant's clothing so that spontaneous cries can be recorded and there is no need to give stimulation for cry generation. Such recording can be helpful in studying the developmental aspects of the infants and for a long-term follow up in the case of sick infants.

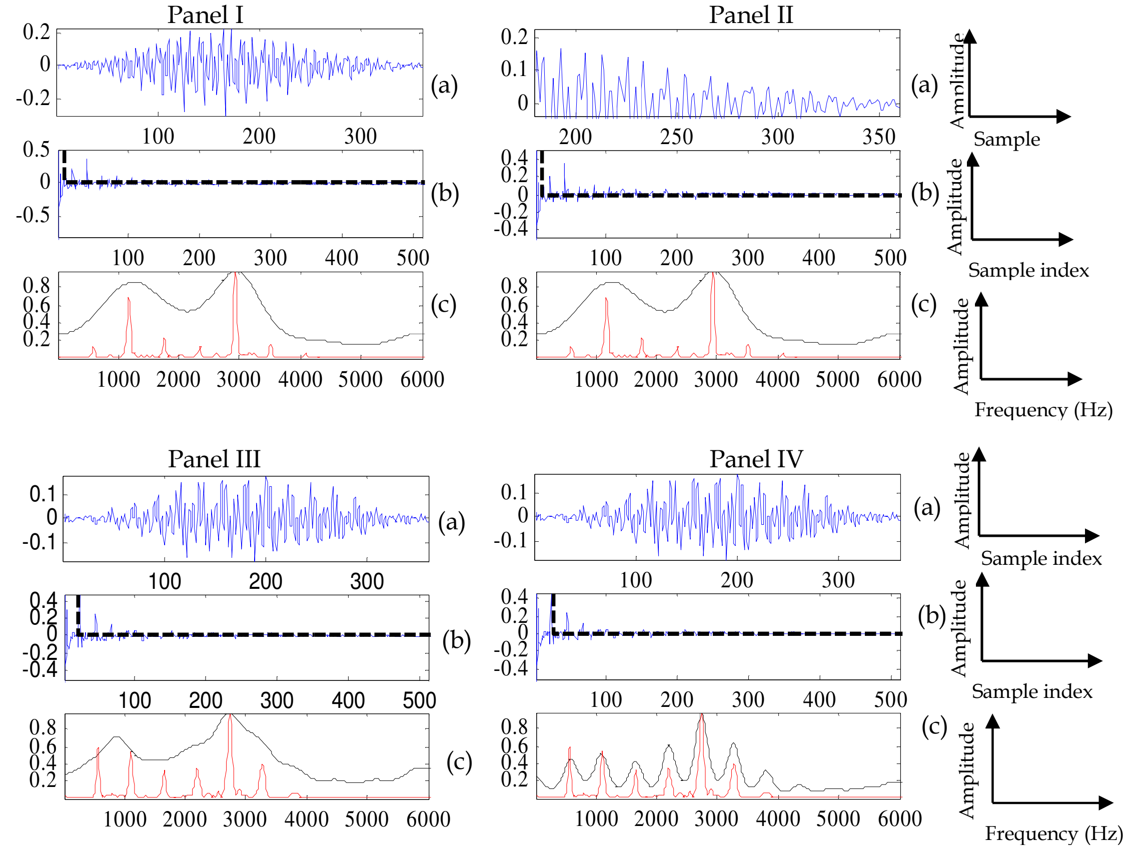


Figure 3.11. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II-*12* samples, Panel III- *20* samples and Panel IV- *25* samples for infant cry signal at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) real cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response superimposed on STFT obtained by liftering (b).

Data should be recorded at a high sampling frequency (F_s) such as *22* kHz or *44.1* kHz. Reasons for using high sampling frequency are as follows:

1. The fundamental frequency (F_0) in infants is higher than the male voice, female voices and children voices (due to the small mass of the vocal folds). F_0 in infants are found in the range of *350* Hz to *1.2* kHz. High sampling frequency, gives better resolution in fundamental frequency estimated using computer-based algorithms.
2. The vocal tract length in infants is small (about *7.5* cm) which is half of the vocal tract length in adults. Small vocal tract length results in high formant frequencies almost double than adult's formants. The theoretical values of formants in infants are *1.1* kHz, *2.2* kHz and so on. Thus, using lower

values of sampling frequency limits the number of formants to be covered in the available spectrum of the infant cry signal which is half of the sampling frequency (due to Shannon's sampling theorem).

3. Choosing a lifter size (*i.e.*, the window in cepstrum-domain) in cepstrum analysis of infants is a difficult task with the smaller sampling frequency. Changing the lifter size by a single sample value in low F_s signal results in large variation in the cepstrum source and system response separation (called as deconvolution). This effect is illustrated in Figure 3.11, using the infant cry sample of 12 kHz (*Corpus I*). It can be observed from the Figure 3.11(a)- Figure 3.11(d), that as the lifter size is changed from *10* samples to *25* samples, the system response capture source information instead of system related information, *i.e.*, formants. Thus, to use cepstrum analysis in an effective way, high sampling frequency is required. Further details of the cepstral analysis of infant cry signal are presented in Section 4.5.
4. Higher F_s provides a sufficient number of samples between two glottal closure instances (GCIs), necessary to detect GCI parameters.

3.9 Chapter Summary

In this chapter, important factors affecting the data collection of infant cries are discussed with suitable references and examples. It is shown that the prosodic marks, duration, voicing, power and short-time energy of the infant cry signal changes significantly with the infant's age, weight and reason of crying. Analysis of infant cries, without considering these variables (*i.e.*, factors) may give misleading experimental results. This study may be helpful for those researchers who want to work in the area of infant cry analysis for the purpose of possible medical diagnosis, developmental studies, infant's/parent's behavioral studies and cry analysis and signal processing research.

Apart from the factors reported in this Chapter, other factors may also affect the cry characteristics such as change of place, crowd, improper handling of the infant, *etc*. though it is difficult to cover all the factors in an analysis and considering more factors in data collection process makes the task even more difficult. Hence, it is upto the research problem and availability of infant cry signals which needs to be balanced by the researcher. Similarly, in the participant information form, other factors can also be included keeping in the mind the possible use of the database for the future research. For example, in the study of effects of a drug on the infant development during the prenatal or postnatal period, one can add a column indicating the use of the drug and its dosage in daily routine or combination of the drugs to find out their possible consequences. In another example, one can also study the effect of multilingual environment in infant cry patterns and can relate it to their language acquisition skills. In all, its upto the vision and area of the researcher to decide which questions or details should be asked while collecting the infant cry samples so that it can be used in a long term and maximum information can be explored from it for particular research problem. After preparation of infant cry corpus, the signal processing challenges associated with infant cry analysis are presented in the next chapter.

Chapter 4.

Signal Processing Challenges in Infant Cry Analysis

4.1 Introduction

In this chapter, signal processing challenges associated with infant cry analysis are presented. Most of the speech signal processing applications, such as, speech recognition, speaker recognition, speech synthesis and voice conversion systems are designed for adult speeches only [**100**], [**101**], [**102**], [**103**], [**104**], [**105**], [**106**], [**107**], [**108**], [**109**]. Some work is done on speech enhancement in noisy conditions for adult voices [**110**], [**111**]. Researchers have worked on the prosodic context of speech for emotion recognition and defining new prosodic features using speech signal and articulators used [**112**], [**113**], [**114**]. Researchers have also made efforts towards understanding the speech production mechanism [**115**], [**116**]. In speech signal processing-domain, female voices are considered difficult to analyze compared to male voices due to their high pitch (or fundamental frequency, F_0) and thus, having associated *spectral resolution* problem [126]. In infants, the pitch is even higher resulting in distantly spaced excitation source harmonics which in turn makes the infant cry signal processing much more difficult and challenging. A comparison of male speech, female speech, child speech and infant cry signal for different signal processing methods is presented in this chapter to illustrate the practical problems associated with applying traditional and state-of-the-art methods directly to the infant cry signals. The analysis presented in this chapter may find its significance in the development of new signal processing algorithms suitable for infant cry analysis and possible technical application of social relevance.

In this chapter, four signal processing methods are explored for infant cry analysis, namely, short-time Fourier transform (STFT) analysis, linear prediction (LP) analysis, cepstral analysis and Teager energy operator (TEO) analysis. In all the analysis, the infant cry samples of *Corpus II* are used. A detailed description of the corpus is given in [**93**]. For the analysis of adult speech, TIMIT database is used which is recorded at a sampling frequency of *18 kHz*. In the TIMIT database, there are sentences spoken by the adult male and female speakers and the duration of these sentences is *1-2* seconds. In all the analysis of speech samples, the sampling frequency of all the signals is downsampled to *12* kHz (if it is higher than this).

4.2 Short-Time Fourier Transform (STFT) Analysis of Speech Signals

The short-time Fourier analysis is widely used in speech analysis applications. Speech is not stationary, especially over a longer duration. Thus, a single Fourier representation of speech does not convey a meaningful interpretation of speech signal. To represent the speech signal as stationary and speech production system to be modeled reasonably as linear and time-invariant (LTI), the speech signal is blocked into short duration overlapping frames of *10-30 ms* duration. On these smaller duration frames, the Fourier analysis is applied. This representation is called Short-Time Fourier Transform (STFT) of a signal. The STFT is mathematically defined for a frame $s[n]$ as:

$$X(m,\omega)=\sum_{n=-\infty}^{\infty} s[n]w[n-m]e^{-j\omega n} =< s(n), w_{n,m}e^{j\omega n} >, \tag{4.1}$$

where $s[n]$ is the signal, $w[n]$ is the window, ω is the frequency and < , > is the *inner product* operator of $s[n]$ with time-frequency atoms $\{w_{n,m}e^{j\omega n}\}$ where $w_{n,m}=w[n-m]$. STFT allows time-frequency analysis of the signal. This representation allows detecting spectral changes with respect to time. In addition, l^2 norm of the speech signal is processed in the STFT-domain, *i.e.*, due to Parseval's energy equivalence theorem, *i.e.*,

$$\sum_{n=0}^{N-1}|s(n)|^2=\frac{1}{n}\sum_{l=0}^{N-1}\sum_{m=0}^{N-1}|X(m,\omega_l)|^2 , \tag{4.2}$$

where $\omega_l=(\frac{2\pi}{N}).l$. In Figure 4.1- Figure 4.4, STFT of the speech segments are shown for the male, female, child and infant sound signals (voiced). In all the figures (Figure 4.1- Figure 4.4), speech utterance of /*aa*/ is taken for analysis. All samples are resampled to *12* kHz for visualizing the four cases on same frequency scale (*i.e.*, to have same available bandwidth due to Shannon's sampling theorem). All the frames are of duration *50 ms* duration (*30* ms in Figure 4.4) with overlapping of *10 ms* and window considered is a rectangular window (for the simplest case). From Figure 4.1 and Figure 4.2, we can observe that the male voice has clear excitation source harmonics structure in its short-time Fourier spectrum. However, the visibility of such harmonics structure and their dominance in the spectrum is not clearly visible in other two cases (*i.e.*, for female and infant). In infants, the harmonics amplitudes are also negligible after fourth harmonics as shown in Figure 4.3.

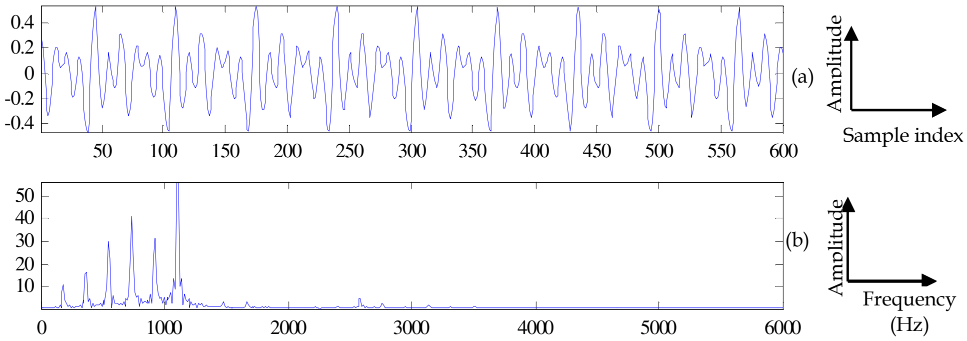


Figure 4.1. Short-time Fourier spectrum of a male speech for vowel /aa/ (a) time-domain signal and (b) corresponding short-time Fourier spectrum at *Fs*= *12* kHz.

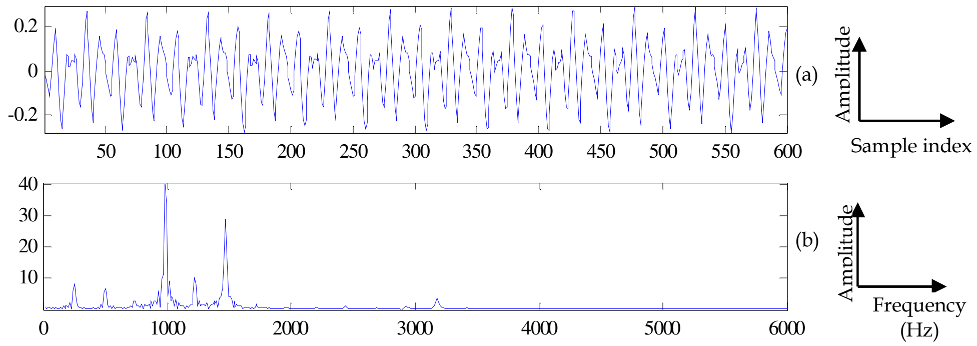


Figure 4.2 Short-time Fourier spectrum of female speech for vowel /*aa*/ (a) time-domain signal of and (b) corresponding short-time Fourier spectrum at *Fs*= *12* kHz.

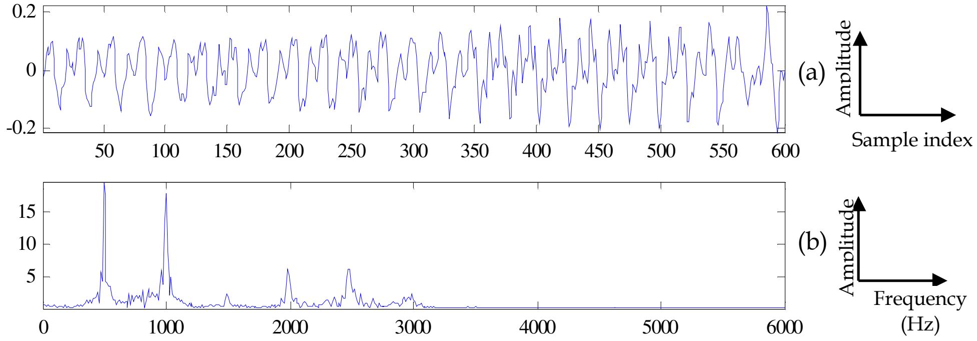


Figure 4.3. Short-time Fourier spectrum of an infant cry (a) time-domain signal and (b) corresponding short-time Fourier spectrum at *Fs*= *12* kHz.

It can also be observed that the fundamental frequency (F_0) is also changing with the age. In adults, once the vocal production system is developed, F_0 also depends on the gender. In the male speech, F_0 ranges around *125* Hz while in the case of a female, it is around *200* Hz spectral range. In children, the values of F_0 are around *250-400* Hz and in infants, it is around *500* Hz range and in some cases, it can raise upto *1* kHz. The differences in F_0 are due to the size of the vocal source (especially, the mass and tension in the vocal folds). The size of the larynx in men is about *40* % taller and longer than the women. The vocal fold length in male speaker is *60* % longer than the female speaker. This reason is responsible for higher F_0 in female speakers. The same reasoning is also valid for children and infants. The relation in F_0 and vocal fold length is given by [**117**]:

$$F_0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}}, \tag{4.3}$$

where L is the length of the vocal folds, σ is the longitudinal stress and ρ is the tissue density in vocal folds. From Figure 4.1 – Figure 4.3, we can observe the differences in the fundamental frequency of the various speakers. Estimation of fundamental frequency from STFT is comparatively easy in male speaker. However, in female speakers and infants, due to higher differences in the amplitudes of the harmonics, the task becomes difficult.

Another difficulty in using Fourier analysis for infant cry signal is in the estimation of formants. The values of the formants for the different speakers corresponding to their different vocal tract lengths are shown in Table 4.1. The formant frequencies are calculated using the formula

$$F_k = \frac{(2k-1)c}{4l}, \tag{4.4}$$

where k is a positive integer ($k = 1,2,3,\ldots$), c is the velocity of sound in air medium (*i.e., 350 m/s*) and l is the length of the vocal tract of the speaker. It can be observed from Table 4.1 that the formants in infants and children are far away from the formants of adult male or female speakers. This is also a reason of considering only lower formants in the analysis of infant and children speeches or voices (since the higher formants are difficult to observe and estimate with the given limited sampling frequency and hence, limited available bandwidth due to Shannon's sampling theorem).

Table 4.1. Variation of formants with speaker age and gender (formant frequencies are in kHz) (the data is not shown for the same subject)

Formants	Male (*l=17 cm*)	Female (*l=14 cm*)	Children (6 yrs.) (*l=11.4 cm*)	Infant (1.5 yrs.) (*l=8.5 cm*)	Infant (at birth) (*l=8 cm*)
F_1	0.514	0.625	0.767	1.029	1.09
F_2	1.5	1.875	2.30	3.088	3.28
F_3	2.5	3.125	3.837	5.147	5.46
F_4	3.6	4.375	5.372	7.205	7.65
F_5	4.6	5.625	6.907	9.264	9.84

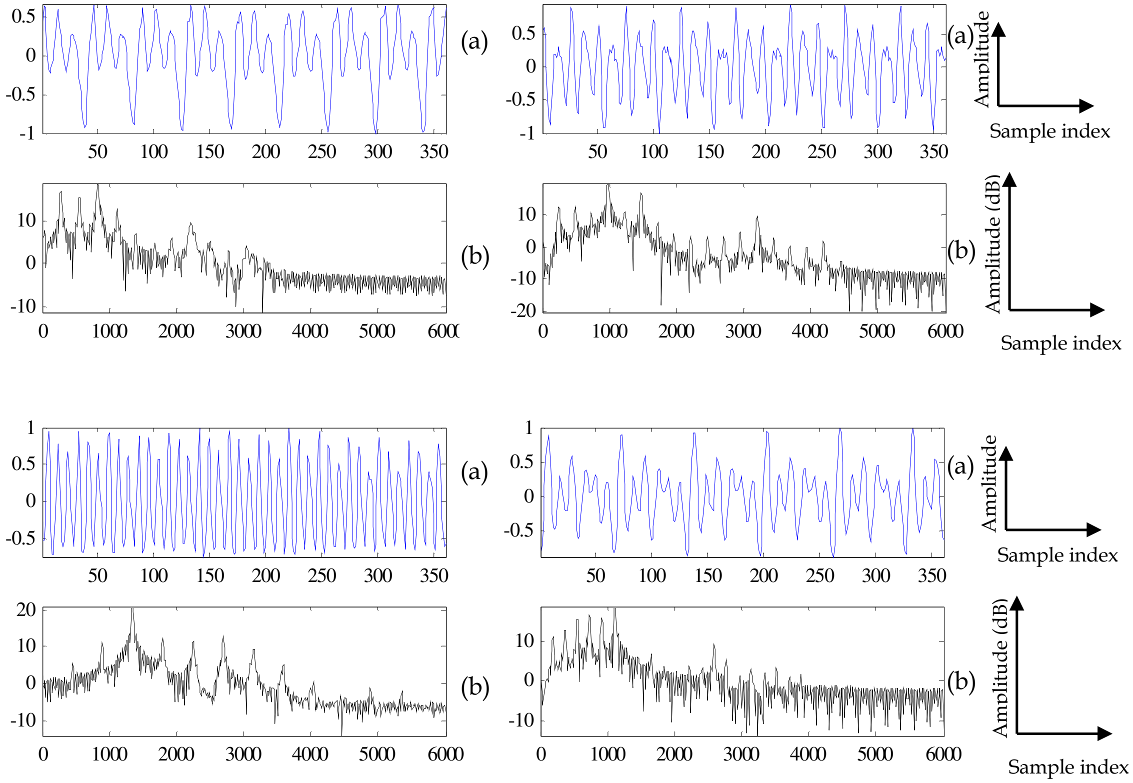


Figure 4.4: Short-time Fourier analysis of male (Panel I), Female (Panel II), Children's speech (Panel III) and infant's cry signal (Panel IV). In all the Figures, (a) shows time-domain signal and (b) shows short-time Fourier spectrum of the signal shown in (a).

In the Fourier spectrum of adult speakers, the peaks of the source harmonics correspond to the fundamental frequency. However, high F_0 in infants results in confusion in formant estimation. To estimate the formants from the STFT spectrum, the log-magnitude spectrum is shown in Figure 4.4. It also shows that higher harmonics also carry a significant role in the spectrum of speech. From Figure 4.4, it is clear that higher harmonics are dominating in the spectrum of high-pitched speakers. Thus, the decision regarding the location of formant frequency cannot be taken from the Fourier spectrum of the signal.

4.3 Linear Prediction (LP) of the Speech

Historically, the idea of linear prediction and all-pole modeling of the system was used in system identification and control literature **[118]**. A model using linear prediction (LP) of speech was first time proposed by Atal and Hanauer [**119**], [**120**]. The model proposed that the speech signal $s(n)$ is produced due to the convolution of impulse-like excitation of excitation source or noise-like source $p(n)$ with the impulse response of vocal tract $h(n)$. The LP model of speech is shown in Figure 4.5 [**121**] [**122**] .

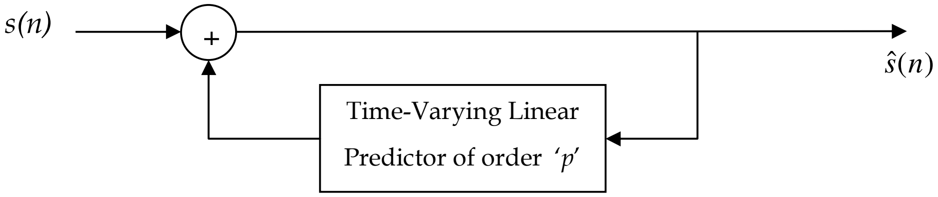


Figure 4.5. All-pole model via Linear Prediction (LP) of the speech signal. After [**119**].

During the production of voiced speech signal, vocal tract system is excited by an impulse-like excitation due to the *sudden* closure of vocal folds and for the production of unvoiced sounds, vocal tract is excited by a noisy excitation source, *i.e.*, turbulent air passing throu+gh the constriction of the vocal tract. In the proposed model, it is assumed that the vocal tract shape remains constant for a small duration of time. Hence, the transfer function of the vocal tract can be approximated from the given relation:

$$s(n) = h(n) * p(n). \tag{4.5}$$

By convolution theorem in Z-domain

$$S(z) = H(z)P(z), \tag{4.6}$$

where $S(z)$ is the Z-transform of the speech segment for a small duration $s(n)$, $H(z)$ is Z- transform of the impulse response of the vocal tract (*i.e.*, $h(n)$) and $P(z)$ is the Z-transform of the excitation source signal, $p(n)$. The transfer function can be represented by its poles and zeros. The unvoiced and nasal sounds contribute to the zeros in the transfer function. However, poles are produced due to the resonance of the vocal tract. For voiced sounds, the transfer function of the vocal tract can be represented as an all-pole model.

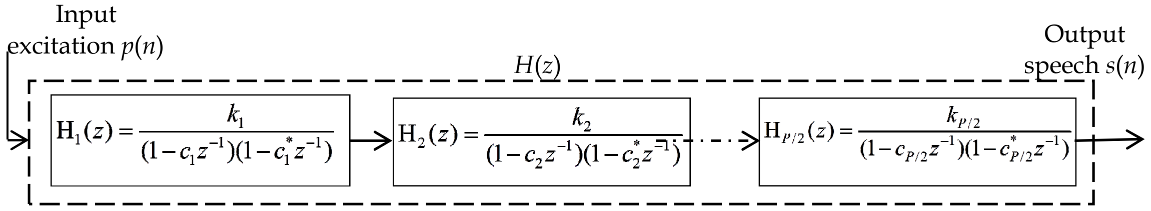

Input
excitation $p(n)$
$H(z)$
Output
speech $s(n)$
$\mathrm{H}_1(z)=\frac{k_1}{(1-c_1z^{-1})(1-c_1^*z^{-1})}$
$\mathrm{H}_2(z)=\frac{k_2}{(1-c_2z^{-1})(1-c_2^*z^{-1})}$
$\mathrm{H}_{P/2}(z)=\frac{k_{P/2}}{(1-c_{P/2}z^{-1})(1-c_{P/2}^*z^{-1})}$

Figure 4.6. Cascade of second order digital resonators for speech generation using LP model.

Speech signal (voiced) at n^{th} instant can be predicted by its previous p samples (where p is LP model order). The estimated speech signal $\hat{s}(n)$, LP prediction error or residual $e(n)$ and Z-domain system fnction of LP moel are given as

$$\hat{s}(n)=\sum_{k=1}^{p}a_k s(n-k), \tag{4.7}$$

$$e(n)=s(n)-\hat{s}(n), \tag{4.8}$$

$$\mathrm{H}(z)=\frac{G}{1-\sum\limits_{k=1}^{p}a_k z^{-k}}=\frac{G}{\prod\limits_{k=1}^{\lfloor p/2\rfloor}(1-c_k z^{-1})(1-c_k^* z^{-1})}, \tag{4.9}$$

where G is the gain factor in LP model (*i.e.*, H(z)) which represents intensity, $e(n)$ is the LP residual or error, $\{a_k\}_{k\in[1,p]}$ are the LP coefficients, $c_k=r_k e^{j\theta_k}$ is the pole of the k^{th} second order resonator (as shown in Figure 4.6) and $a_k{}'$s are computed by using the autocorrelation method. Since, $H(z)$ contains only poles and in principle, no zeros, $H(z)$ is called as an *all-pole* model in system identification and signal processing literature. The magnitude response of the vocal tract system is given by the LP spectrum which is obtained as

$$|\mathrm{H}(e^{j\omega})|=\left|\frac{G}{1-\sum\limits_{k=1}^{p}a_k e^{-j\omega k}}\right|. \tag{4.10}$$

The resonances of the vocal tract system are called as formant frequencies. The term $|\mathrm{H}(e^{j\omega})|^2$ is called as linear prediction (LP) spectrum or LP model spectrum. To get better spectral resolution in LP spectrum computation, zero-padding is used. The LP analysis has been used successfully for many years in speech analysis and synthesis. The LP order, *p,* is dependent on the vocal tract length. To model the vocal tract filter, the memory of the vocal tract filter must be at least *twice* the time required for the sound wave to travel from glottis to the lips [**119**]. This time interval is given as $\tau = 2l / c$, where $c = 350\, m/s$ is the speed of sounds in air medium. To account for lip radiation effect and effect of the nasal cavity, *4* additional poles are added in the estimated prediction order *p*. The length of vocal tract for a male, female speaker and newborn is *17 cm*, *14 cm* and *8 cm,* respectively [**26**]. The estimated prediction coefficients are *16*, *14* and *10* for an adult male, female and newborns, respectively. In the next Section, the effect of LP order in signal processing is discussed for male, female and infant.

4.4 Analysis of Results with respect to Spectral Matching

For voiced speech, excitation source is impulse-like sequence of impulses. Hence, its Fourier spectrum due to spectral matching LP problem formulation, *i.e.*, in *Z*- domain, we have,

$$e(n) = s(n) - \hat{s}(n), \tag{4.11}$$

$$Z\{\mathrm{e(n)}\} = \mathrm{Z}\{\mathrm{s(n)}\} - \mathrm{Z}\{\hat{\mathrm{s}}\mathrm{(n)}\}, \tag{4.12}$$

$$E(z) = S(z)[1 - \sum_{k=1}^{p} a_k z^{-k}] = \mathrm{S(z)\,A(z)}, \tag{4.13}$$

$$E(z) = \frac{S(z)}{H(z)}, \tag{4.14}$$

$$E(e^{j\omega}) = \frac{S(e^{j\omega})}{H(e^{j\omega})}. \tag{4.15}$$

The l^2 norm of LP residual is preserved, *i.e.*, by Parseval's energy equivalence for discrete-time Fourier transform (DTFT) [121],

$$\sum_{n=0}^{N-1} | \mathrm{e}(\mathrm{n}) |^2 = \frac{1}{2\pi} \int_0^{2\pi} | E(e^{j\omega}) |^2 d\omega \ , \tag{4.16}$$

$$\sum_{n=0}^{N-1} | e(n) |^2 = \frac{1}{2\pi} \int_0^{2\pi} \frac{| \mathrm{S}(\mathrm{e}^{j\omega}) |^2}{| \mathrm{H}(\mathrm{e}^{j\omega}) |^2} d\omega = \frac{1}{2\pi} \int_0^{2\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega, \tag{4.17}$$

where $P(\omega)$ is signal spectrum (corresponds to $| S(e^{j\omega}) |^2$) and $\hat{P}(\omega)$ is the LP spectrum (corresponds to $| H(e^{j\omega}) |^2$). While estimating optimum values of LP coefficients (*i.e.*, $\{\mathrm{a}_k\}_{k \in \lceil 1,p \rceil}$), our goal is to minimize l^2 norm of LP residual and hence, to minimize $\frac{P(\omega)}{\hat{P}(\omega)}$ in the above integral [121]. In doing so, the LP model spectrum, *i.e.*, $\hat{\mathrm{P}}(\omega)$ tries to match the first dominant peak in the STFT spectrum (which is nothing but the formant frequency of vocal tract) and then LP spectrum matches to second dominant peak whose height is less than that of first formant due to various energy losses such as wall vibrations, lip radiation, thermal conductance, *etc.* during speech production [**98**].

4.4.1 Effect of Order of Linear Predictor

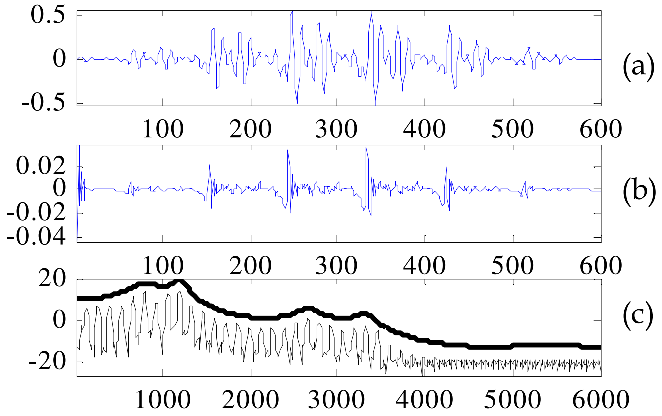


Figure 4.7. LP analysis of a male speech for vowel */aa/* (a) time-domain signal (b) corresponding LP residual and (c) corresponding LP spectrum for LP order of *12* shown by dark line superimposed on STFT .

The LP analysis of a male speech is shown in Figure 4.7. The speech is first converted into smaller frames of *50 ms*, because the LP model considers the signal under consideration as output of linear time invariant (LTI) system.

In Figure 4.7 (a), the time-domain signal is shown. From this time-domain signal, its estimated signal is evaluated using eq. (4.7) with LP order of $p=12$. Then LP error or residual signal is estimated by taking the difference of the original speech segment and estimated speech segment (using LP) which is plotted in subfigure Figure 4.7 (b). Then, the log-magnitude spectrum of the LP residual and the log-magnitude of short-time FFT spectrum (in dB) are plotted in subfigure Figure 4.7 (c). The resonance peaks of the both spectra match when the order of the LP analysis is low (as suggested in Section 4.3). Similar analysis is repeated for male, female, children speech and infant cry in this Section for various LP orders. In Figure 4.8 – Figure 4.11, the subfigure (c) shows the LP spectrum. The legend is the same as that used in Figure 4.7. The light solid line represents the STFT spectrum and the dark solid line represents the LP spectrum.

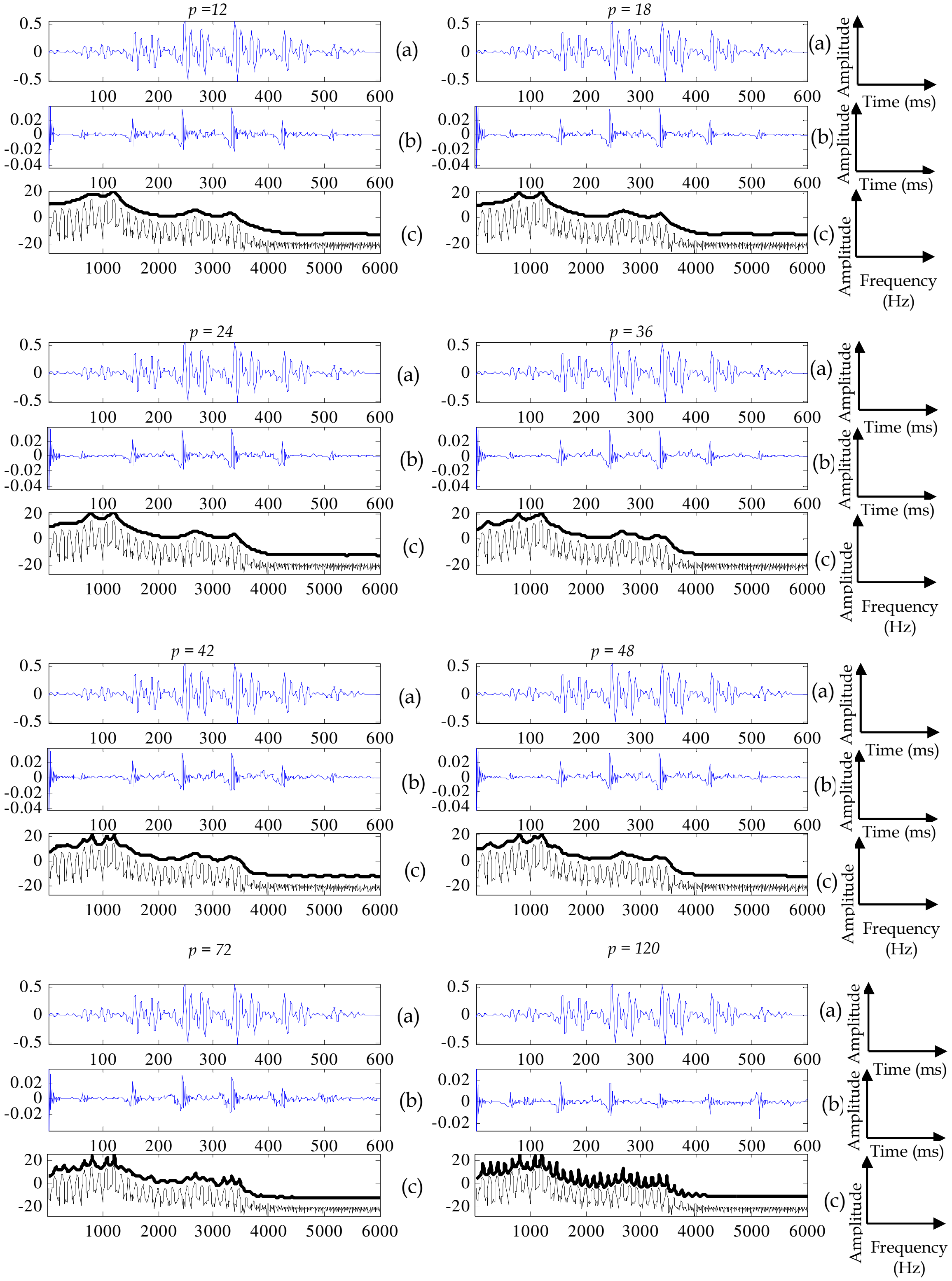


Figure 4.8. Effect of LP order p in male speech analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures, F_s=*12* kHz. In each subfigure (c), the amplitude is in dB.

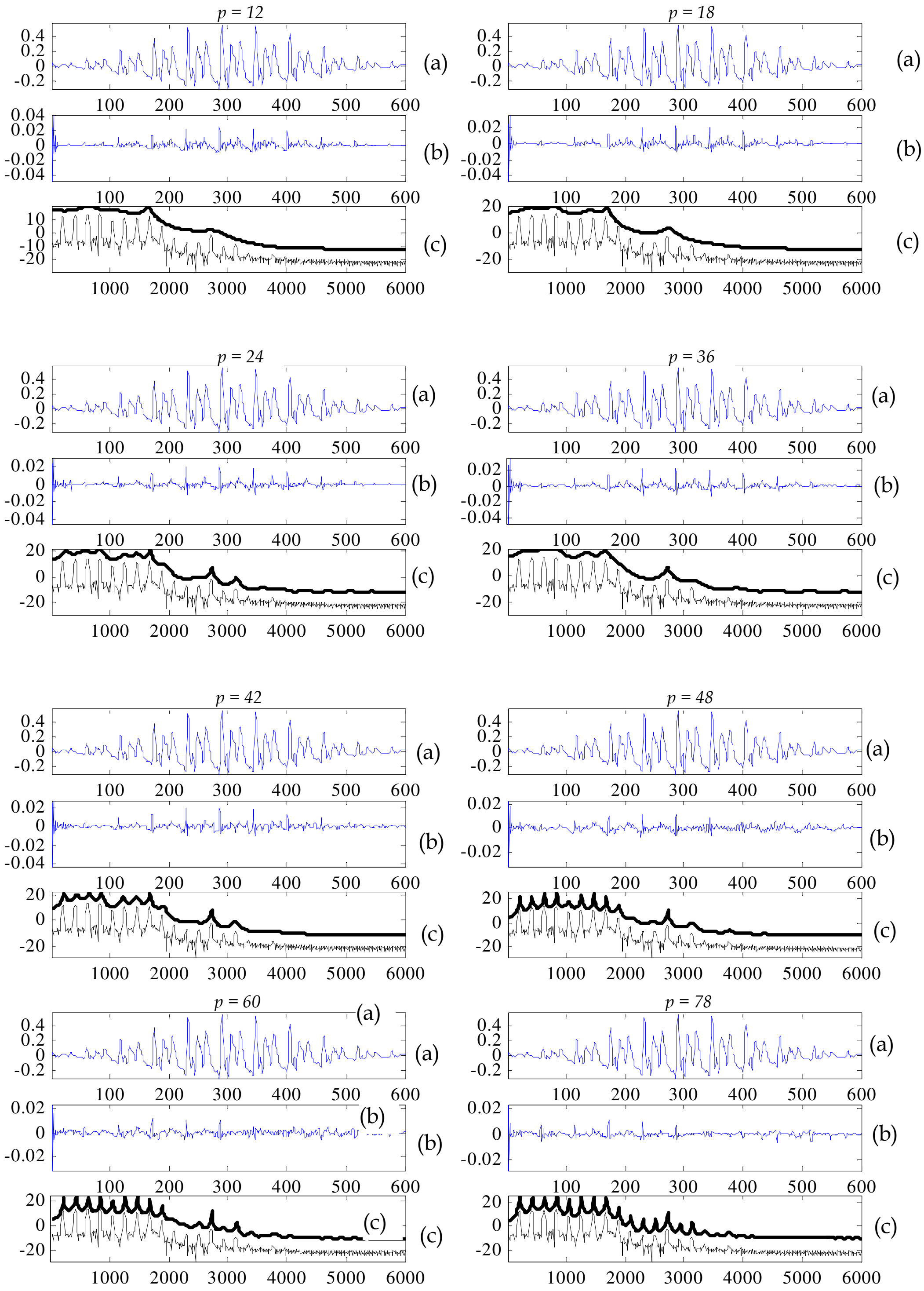
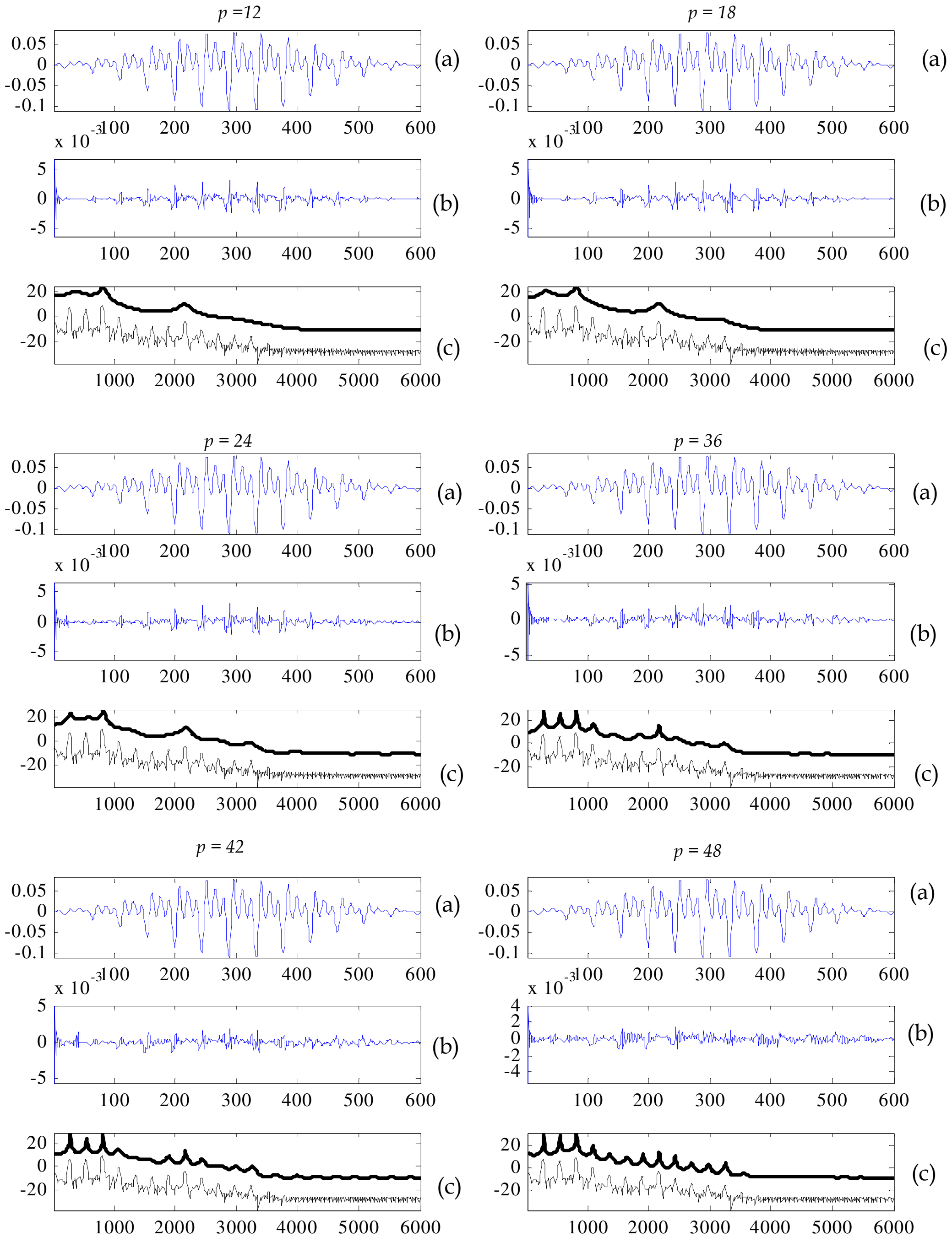


Figure 4.9. Effect of LP order p in female speech analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures: (a) and (b) X- axis is sample index and in subfigure (c) the X-axis is the frequency in Hz. In all the subfigures, Y-axis is amplitude, F_s=12 kHz. In each subfigure (c), the amplitude is in dB.


p =12

p = 18

p = 24

p = 36

p = 42

p = 48

Figure 4.10 Effect of LP order p in child speech analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures: (a) and (b) X-axis is sample index and in subfigure (c) the X-axis is the frequency in Hz. In all the subfigures, Y-axis is amplitude and F_s=*12* kHz. In each subfigure (c), the amplitude is in dB.

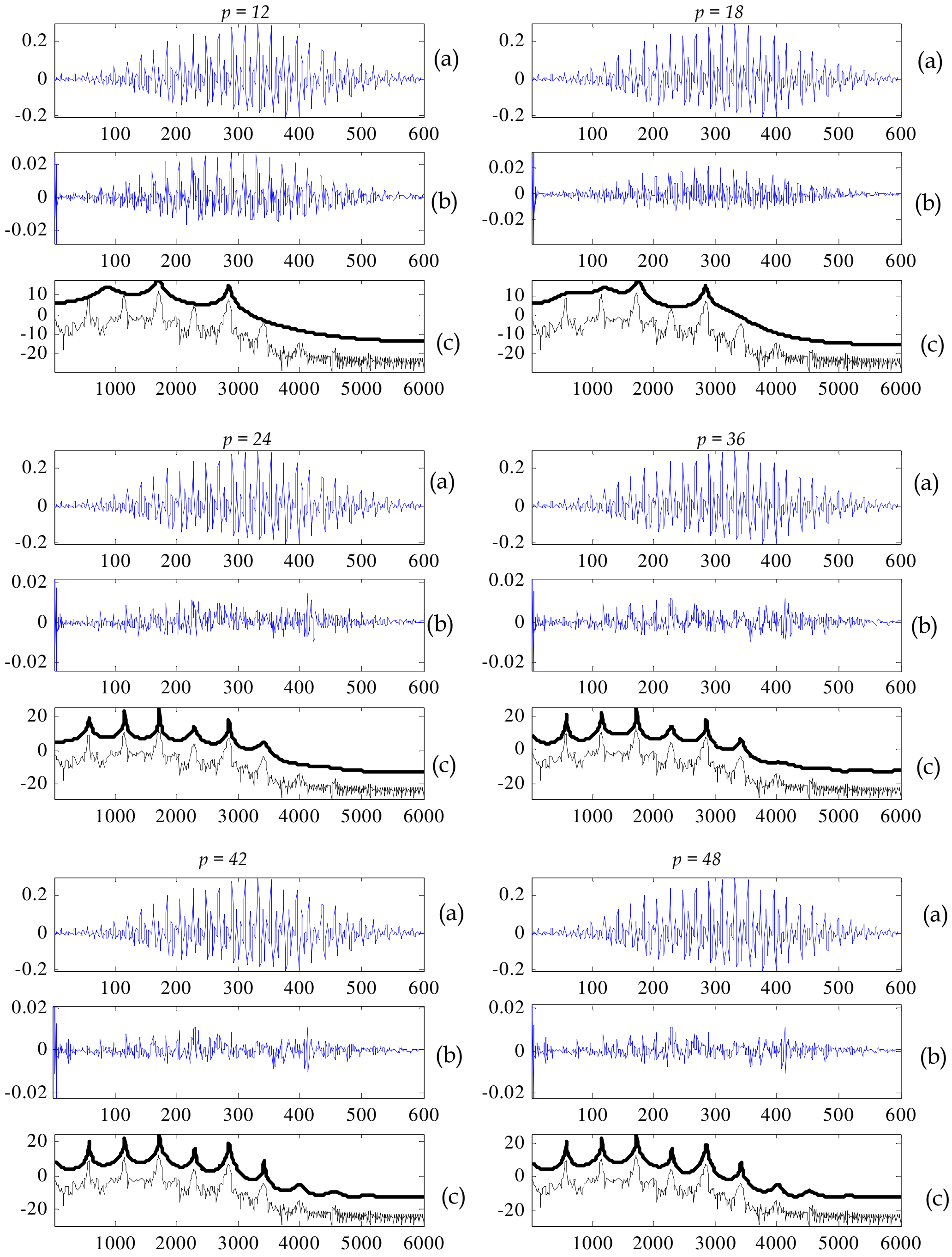


Figure 4.11. Effect of LP order p in the infant cry signal analysis. (a) time-domain signal (b) corresponding LP error signal and (c) LP and short-time Fourier spectra. In all subfigures: (a) and (b) X- axis is samples and in subfigure (c) the X-axis is the frequency in Hz. In all the subfigures, Y-axis is amplitude and F_s=*12* kHz. In each subfigure (c), the amplitude is in dB.

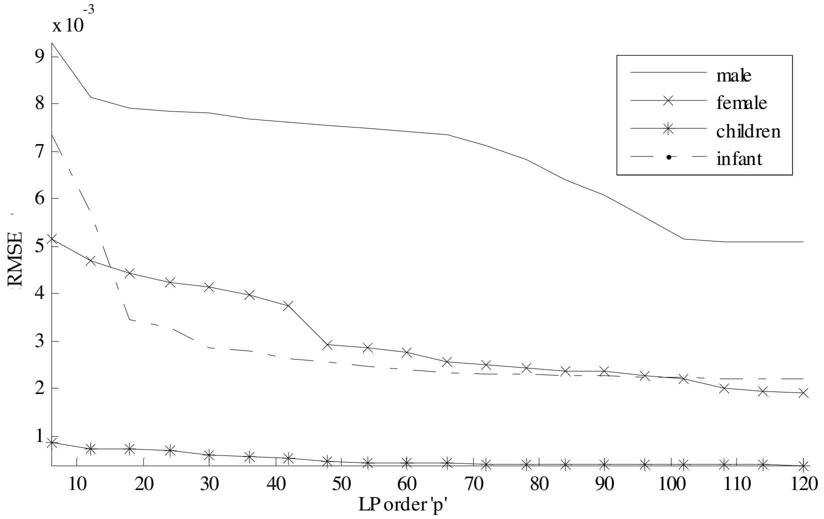


Figure 4.12. LP order *vs*. LP root mean square (rms) error.

Figure 4.8 – Figure 4.11 shows the LP analysis of the voiced speech segments of male, female, child speaker and infant cry signal. It can be observed from Figure 4.8 – Figure 4.11 that with increasing order of LP, the LP spectra matches the peaks of the pitch source harmonics. The value of p for which this happens is *108* for male, *60* for female, *48* for the child and *24* for an infant. However, for formant analysis, required p values are *16*, *14*, *12* and *10* for male, female, children (6 years old) and infants, respectively, for a sampling frequency of *12* kHz. Another observation from all the plots is that increasing the order of LP analysis results in decrease in energy of LP residual or error (*i.e.,* its l^2 norm) as shown in Figure 4.12. In addition to this effect, the LP residual (error) cannot be used for pitch estimation for higher values of p (as it creates more ambiguity in the location of GCIs due to bipolar nature of LP residual). The above mentioned effects are due to the modeling of the vocal tract using LP model, which is dominated by pole (predominantly complex conjugate pole pairs) structure as shown in eq. (4.9). Increasing the order of LP, increases the number of poles in the vocal tract response, and these poles represent the resonances of the vocal tract. A higher number of resonance peaks then corresponds to the excitation source harmonics (F_0) instead of system harmonics (*i.e.,* formants). Moreover, increasing the LP

order p, takes into account more number of samples of the speech signal during prediction (due to eq. (4.9)) and hence, results in more accurate modeling of the speech signal approximation, thereby, reduces the LP error. Interestingly, in Figure 4.12, it can be observed that the l^2 norm of the residual exhibit a sharp decay initially (upto LP order *10-15*) indicating that the all-pole LP model first tries to match the dominant peaks in the spectrum (which corresponds to the formants). Afterwards, there is little gradual decay in l^2 norm and then, it remains almost constant indicating no more optimization of LPCs is possible. This in turn means that speech samples are also related to each other with dependence which is *nonlinear* in nature [**123**]. The gradual decay in the l^2 norm indicates that the LP spectrum tries to match the other dominant peaks (except formants) in STFT (these peaks are nothing but peaks due to pitch source harmonics). Interestingly, this spectral matching happens at different LP orders for different speakers (male-to-infant). Due to sampling of vocal tract spectrum by the very distantly-spaced source harmonics, the spectral peaks in the STFT of infant cry are of almost similar height and hence, LP model tries to match all the peaks simultaneously for comparatively larger values of LP order than that of children, female and male speakers.

From Figure 4.8 - Figure 4.11, it can be seen that as the order of LP analysis is increased, the LP spectrum matches the STFT spectrum and the first formant frequency approaches to the fundamental frequency of the vocal folds vibration. From the plot of the variation of formants with p, in the case of infants (shown in Figure 4.13), it can be observed that with the same experimental setup, it is difficult to find the formants as well. Even with small values of p, harmonics of the fundamental frequency (F_0) are detected and hence, it is difficult to estimate formants automatically with LP analysis. This happens because of the addition of more poles in eq. (4.9) and interaction of excitation source harmonics (which also represents localized dominant peaks) with the vocal tract spectrum. Due to small and narrow vocal tract cross-

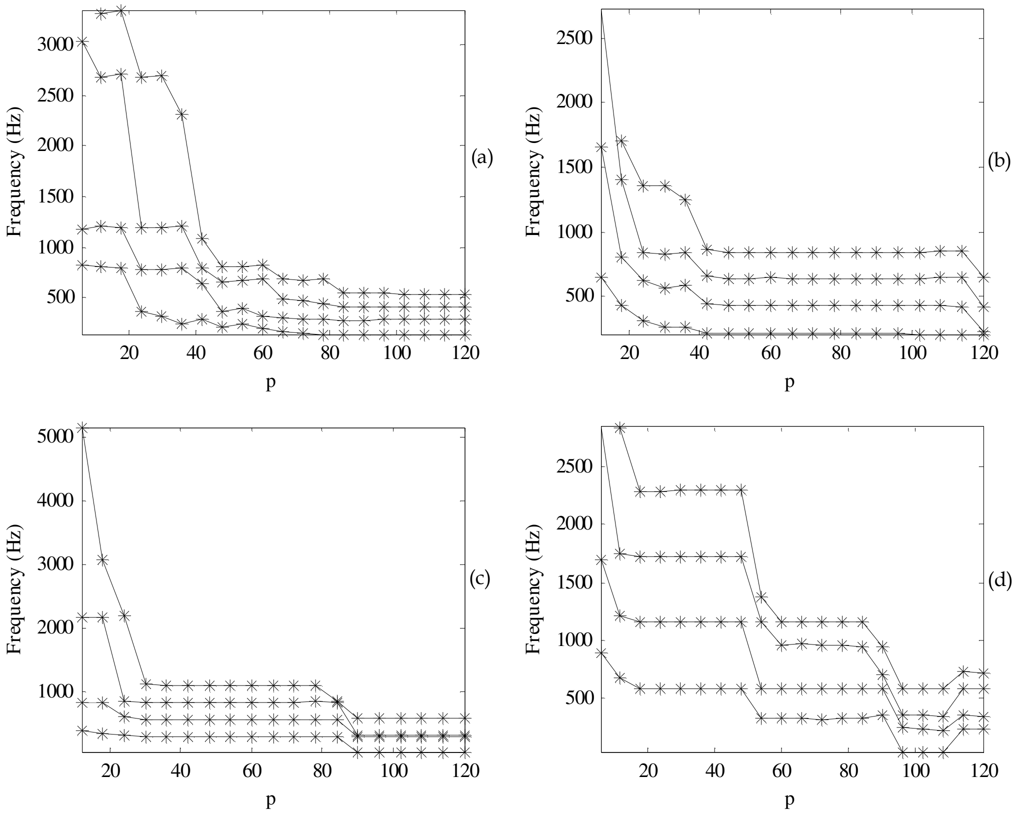


Figure 4.13. Transition of formants into F_0 harmonics with increasing order of linear predictor (LP). (a) male speech, (b) female speech, (c) children speech and (d) infant cry signal.

section, estimation of formants in infants is even more challenging task compared to female and child speech. Along with this, due to the light weight of vocal folds and sometimes underdeveloped vocal folds or larynx, estimation of the fundamental frequency (F_0) is even more difficult task. Thus, well known spectral resolution problem associated with female speech becomes extremely *severe* with associated spectrum of infant cry signal [**124**], [**125**], [**126**] . Moreover, it is clear from the Figure 4.13 that for the same sampling frequency (F_S), the number of formants covered in the range [0, $F_S/2$], is almost half in infants compared to the adults (because of the fact that vocal tract length is almost half in infants compared to the adult speakers due to eq. (4.4)). This draws an important observation that in the case of adult speech analysis, sampling frequency (F_S) as low as *12 kHz* (or even *8* kHz) is sufficient. However, for infant cry, to extract system-related information, the sampling frequency should be kept especially high to capture more formants.

4.5 Cepstral Analysis on Infant Cry Signal

Voiced speech signal is modeled as the response of the vocal tract to the impulse-like excitation (due to the *sudden* closure of glottis) which is produced by the vocal source (*i.e.*, the vocal folds). The speech signal $s(n)$ can be represented in the signal processing framework as convolution of $h(n)$ with $p(n)$ (*i.e.,* eq. (4.5)): The real cepstrum of a signal is obtained by taking the inverse Fourier transform of the logarithm of the Fourier spectrum of the signal, *i.e.*, $IFFT(log(|FFT(\mathrm{s(n)})|))$ (*i.e.*, Fourier transform phase is ignored). Application of logarithm on the magnitude of the signal spectrum makes the convolution operation additive in frequency-domain, which facilitates the deconvolution of excitation source and vocal tract system responses, *i.e.*,

$$S(\omega) = H(\omega).P(\omega), \tag{4.18}$$

where $S(\omega) = |S(\omega)|e^{j\measuredangle S(\omega)}$, $H(\omega) = |H(\omega)|e^{j\measuredangle H(\omega)}$ and $P(\omega) = |P(\omega)|e^{j\measuredangle P(\omega)}$. Hence, $S(\omega) = H(\omega).P(\omega)$ and $\measuredangle S(\omega) = \measuredangle H(\omega) + \measuredangle P(\omega)$.

$$\log(|s(\omega)|) = \log(|\mathrm{H}(\omega)|) + \log(|P(\omega)|). \tag{4.19}$$

$$\hat{S}(\omega) = \hat{H}(\omega) + \hat{P}(\omega)\,, \tag{4.20}$$

Applying inverse Fourier transform to the eq. (4.20) gives,

$$\hat{s}(n) = \hat{h}(\mathrm{n}) + \hat{p}(n), \tag{4.21}$$

where $\hat{s}(n)$ is the real cepstrum of $s(n)$, $\hat{h}(n)$ is the real cepstrum of $h(n)$ and $\hat{p}(n)$ is the real cepstrum of $p(n)$ (since Fourier transform phase of the signal is ignored). Thus, the excitation source and vocal tract system response can be deconvolved. Deconvolution using *lifter* $\{l(n)\}$ (which is linear frequency invariant (LFI) filter) results in flattening or blunting of the spectral peaks corresponding to the formant frequencies (which happen primarily due to logarithm operation and convolution) [**125**].

$$\hat{s}(n).\text{l(n)} = \hat{h}(n).\text{l(n)}, \text{ (ignoring pitch harmonics)} \tag{4.22}$$

$$DFT\{\hat{s}(n).l(n)\} = \hat{H}(\omega) * L(\omega) = L(\omega) * \log\{H(\omega)\}, \tag{4.23}$$

$$\therefore \text{ Cepstrally smoothed spectrum } = L(\omega) * \log(H(\omega)) \tag{4.24}$$

Due to convolution and logarithm operation smoothing of spectral envelope takes place and hence, formant peaks are *blunted* [**125**]. In the voiced sounds, the excitation term corresponds to an event that is relatively extended in time (*i.e.*, pulse-train with pulses every *5-10* ms) and thus, it yields a spectrum that is characterized by a relatively fast varying function of ω, in comparison because of the relatively short impulse response of the vocal tract, its spectrum varies more slowly with ω. Thus, the first part of the cepstrum for the voiced speech corresponds to the cepstrum of the vocal tract impulse response [**125**]. In the cepstrum, the low frequency ripples (*i.e.,* quefreqncy) corresponds to vocal tract response and the high frequency ripples represents the harmonics of the vocal fold's response. Thus, the two responses can be separated by using an appropriate *lifter* (LTI filter term appear in frequency-domain. However, the term lifter does the same job of separating a band of frequency, however, in time-domain or quefrency-domain) and hence, the lifter is referred to as LFI filter.

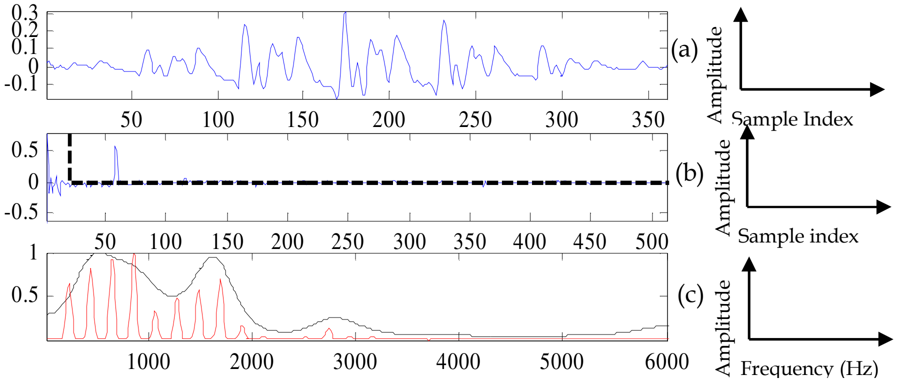


Figure 4.14. Cepstrum analysis of a speech segment of female speaker (a) time-domain signal, (b) real cepstrum of (a) and (c) estimated cepstrally smoothed vocal tract frequency response with lifter size *20* samples superimposed on STFT.

Cepstrum analysis for a short speech segment of *30 ms* taken from the female speech of vowel /*aa*/ is shown in Figure 4.14. In the Figure 4.14 (a), time-domain segment of speech is shown and its real cepstrum is plotted in Figure 4.14 (b). Cepstrum of the signal has vocal tract and excitation source responses. To extract the system response, the initial portion of the cepstrum is separated from the cepstrum using a lifter in time-domain (shown by the thick dotted line in Figure 4.14 (b)). Then, the liftered signal is Fourier transformed to get the vocal tract response in the frequency-domain. The peaks of the vocal tract response (shown in Figure 4.14 (c)) correspond to cepstrally smoothed vocal tract frequency response superimposed on STFT, *i.e.*, formants. In each subfigure (c) of Figure 4.14 - Figure 4.18, thick line corresponds to the cepstrally smoothed vocal tract response and the thin line represents the STFT of the short-time signal shown in each subfigure (a). Similar analysis is performed for all speech segments of various speakers.

The cepstral analysis applied to the male, female, child and infant voices are shown in Figure 4.15 – Figure 4.18. For a speech frame of *30 ms*, the cepstral analysis is shown in Figure 4.15 – Figure 4.18. It can be observed from these figures that cepstral analysis performs better in capturing the formant locations in all the speakers, thereby removing the limitation of LP being not able to capture formant information in high-pitched speakers (especially children and infants). To get formant information from the cepstrum of the speech signal, the size of lifter should be chosen small enough to keep the system-related information and ignore the source-related information in order to minimize source interference in vocal tract system’s frequency response.

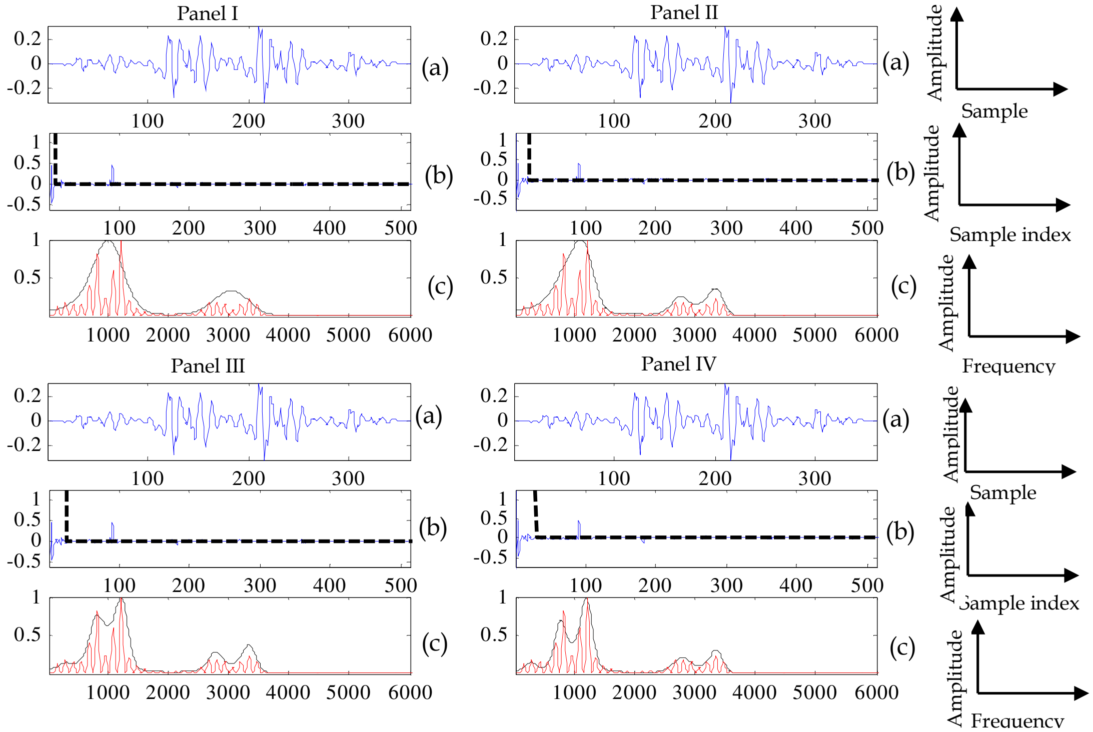


Figure 4.15. Cepstrum analysis for a male speech at the sampling frequency of *12* kHz with varying lifter sizes. The lifter sizes are: Panel I- *10* samples, Panel II-*20* samples , Panel III- *25* samples and Panel IV- *30* samples. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT.

To find the impulse response of vocal tract, the lifter size in case of infants is as low as *10-15* samples is sufficient while increasing the lifter size to *30* samples (required lifter size in male voice for a vowel with sampling frequency of *12* kHz) results in detection of excitation source harmonics in place of impulse response of vocal tract. The reason is the high pitch (and hence, less pitch period of *1-2 ms*) in infant cries which results in serious interference of pitch source harmonics with the impulse response of vocal tract. Thus, it is very difficult to decide the boundary between cepstrum of the excitation source and vocal tract system impulse response. On the other hand, such a small lifter length is not sufficient to capture formant information in adult voices. Another important point to note here is that in the case of adult speakers, the same lifter length can be applied to the analysis of speech signal of different speakers. However, in the case of infants, the lifter length cannot be set to a constant value because of the variability of pitch (F_0) values in infants of different ages and weights. Thus, it requires development of an algorithm to detect the optimum lifter size for analysis of infant cry using cepstrum analysis.

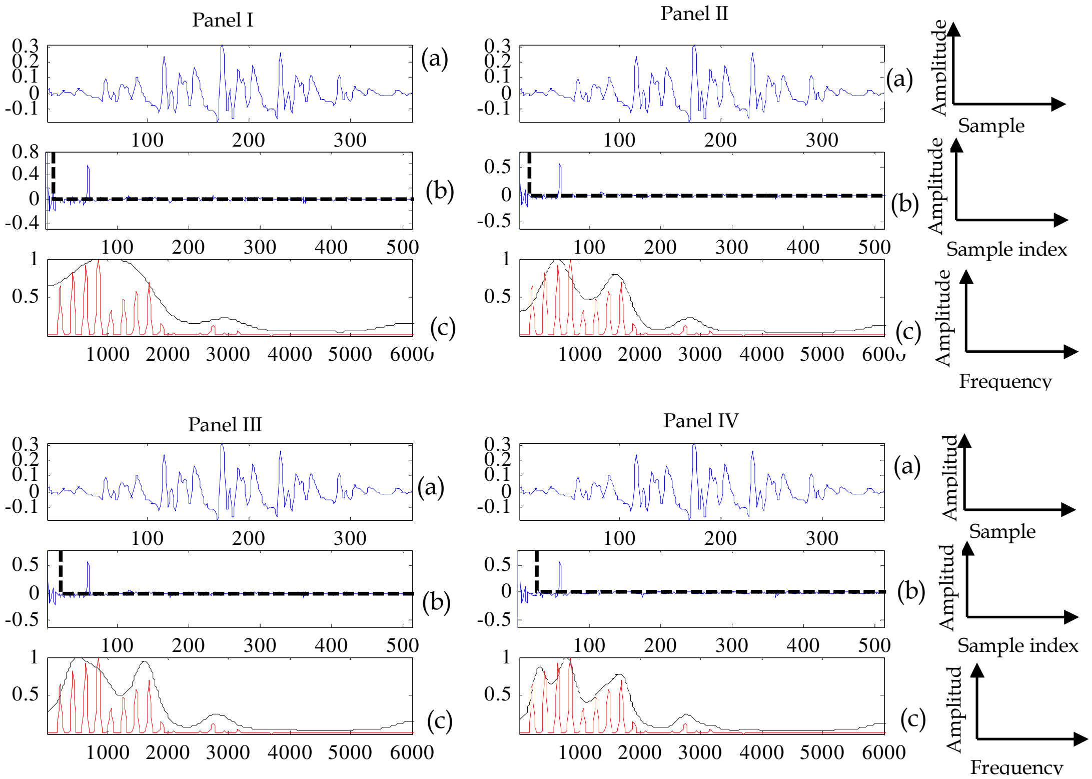


Figure 4.16. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II- *15* samples , Panel III- *20* samples and Panel IV- *25* samples for a female speech at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT.

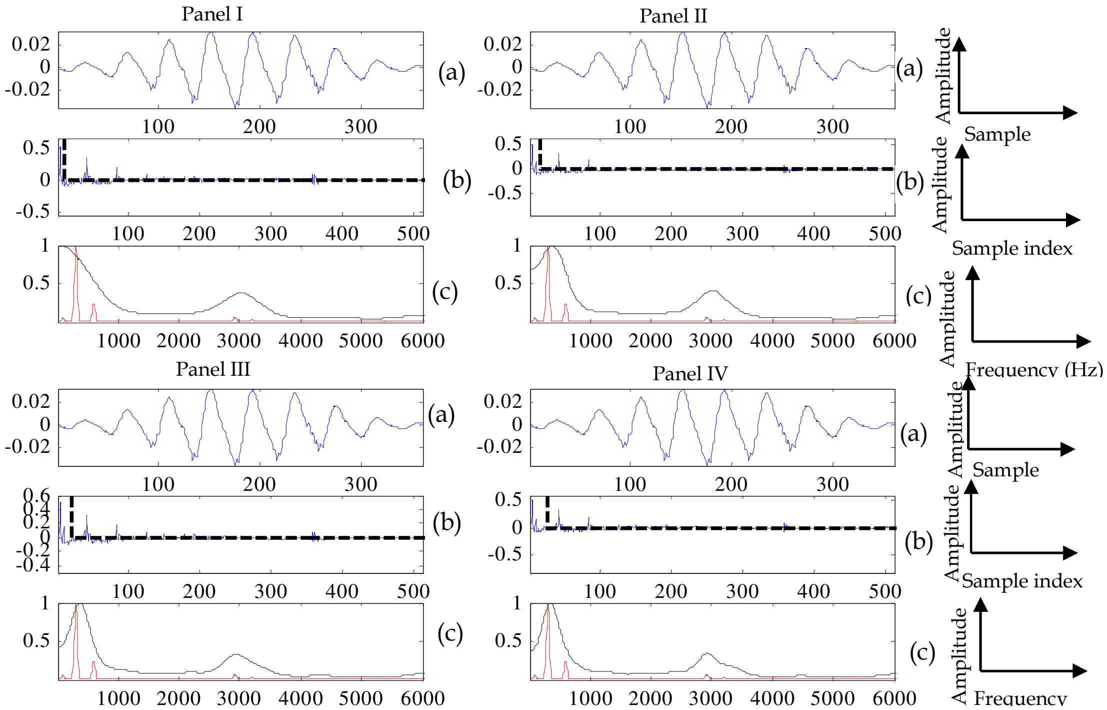


Figure 4.17. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II-*15* samples, Panel III- *20* samples and Panel IV- *25* samples for child speech at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT.

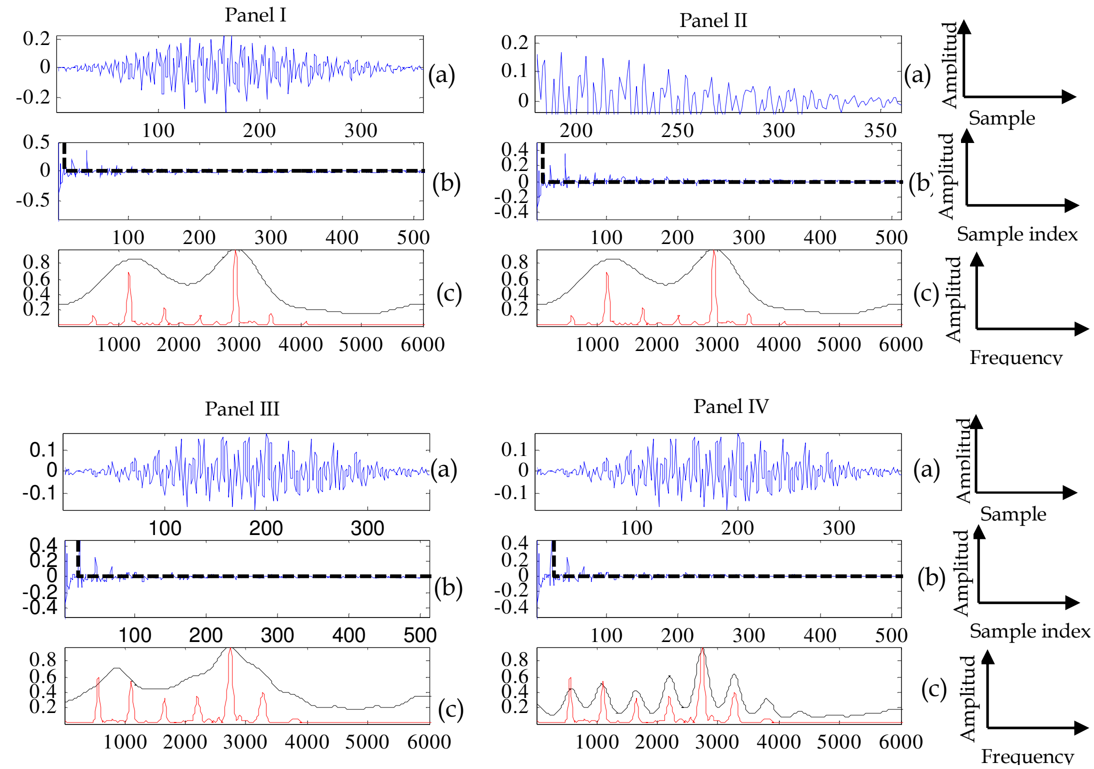


Figure 4.18. Cepstrum analysis with lifter size Panel I- *10* samples, Panel II-*12* samples, Panel III- *20* samples and Panel IV- *25* samples for infant cry signal at the sampling frequency of *12* kHz. In all the subfigures: (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response superimposed on STFT obtained by liftering (b).

4.6 TEO Analysis of the Infant Cry Signal

The methods used in conventional signal processing assume the speech production system as an LTI system. However, the actual speech production system is nonlinear and it is discussed in this Section. According to the Teagers, the pulsatile (for voiced speech) and random noise (for aperiodic speech) airflow which is the excitation source of speech production is not separate and concomitant vortices are distributed throughout the vocal tract. They suggested that the actual source of speech production is vortex and airflow interactions, which are nonlinear. Thus, Teagers suggested a nonlinear model of speech production using the energy of the airflow. This model suggests an energy tracking operator known as Teager Energy Operator (TEO) [**127**], [**128**]. For a short speech segment $x(n)$, TEO is given by

$$\psi\{x(n)\} = x^2(n) - x(n-1).x(n+1), \tag{4.25}$$

where $\psi\{.\}$ denotes the TEO operator. It is observed from the eq. (4.25) that the TEO is dependent only on the previous, present and next sample of the signal (and thus, high time resolution) and gives the running estimate of the signal's energy which may have positive or negative polarity. In this Section, for different speakers (such as male, female, children and infant), TEO operator is analyzed and the differences and similarities among them are identified. For the TEO analysis of speech signals of various speakers, the sampling frequency is kept constant at *12* kHz by resampling the speech segments. The speech signal is lowpass filtered to *5* kHz after resampling and then segmented into the frames of *50 ms* with *10 ms* overlap. In all the analysis, the LP residual $e(n)$ (*i.e.,* eq.(4.11) is also plotted to compare it with the TEO profile of the signal under consideration.

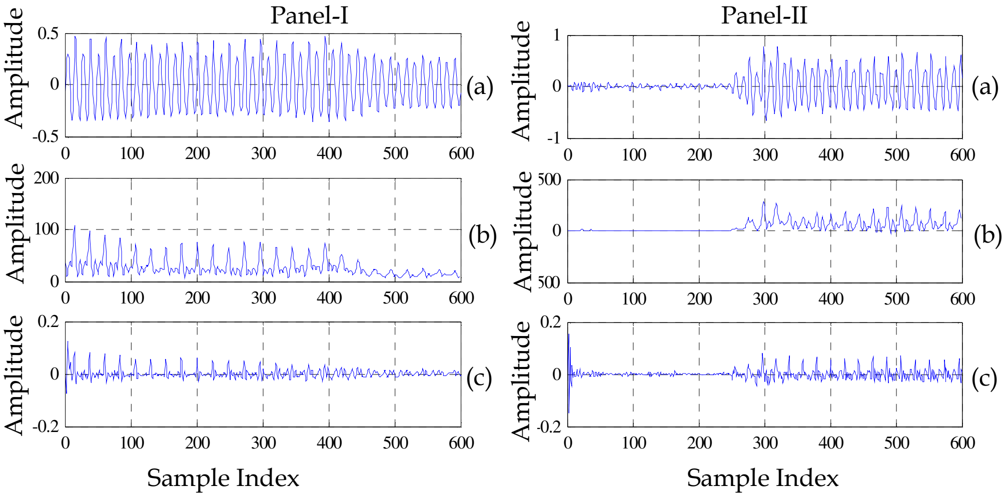


Figure 4.19. Panel I- Voiced infant cry signal and Panel II- Voiced infant cry signal preceded by a silence. In all subfigures: (a) Time-domain signal of an infant's cry and (b) corresponding TEO profile and (c) LP residual of (a).

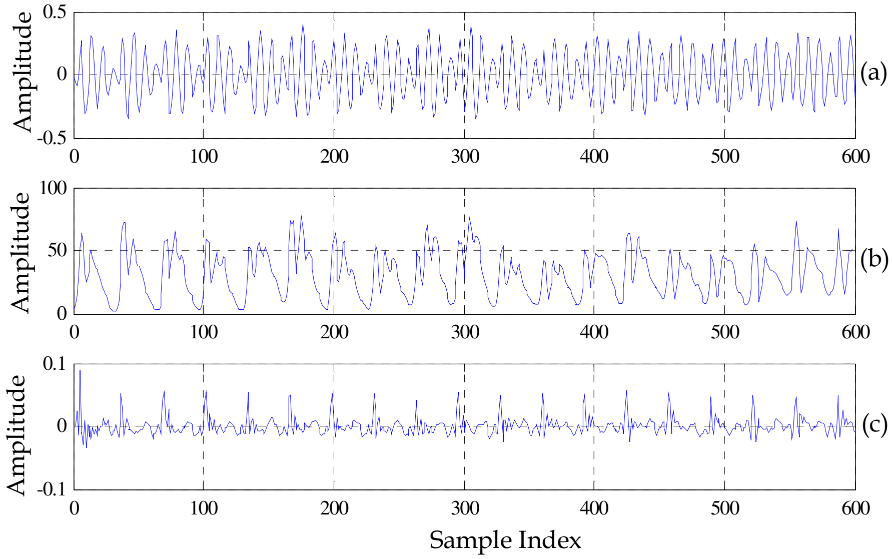


Figure 4.20. TEO analysis of voiced normal infant cry signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a).

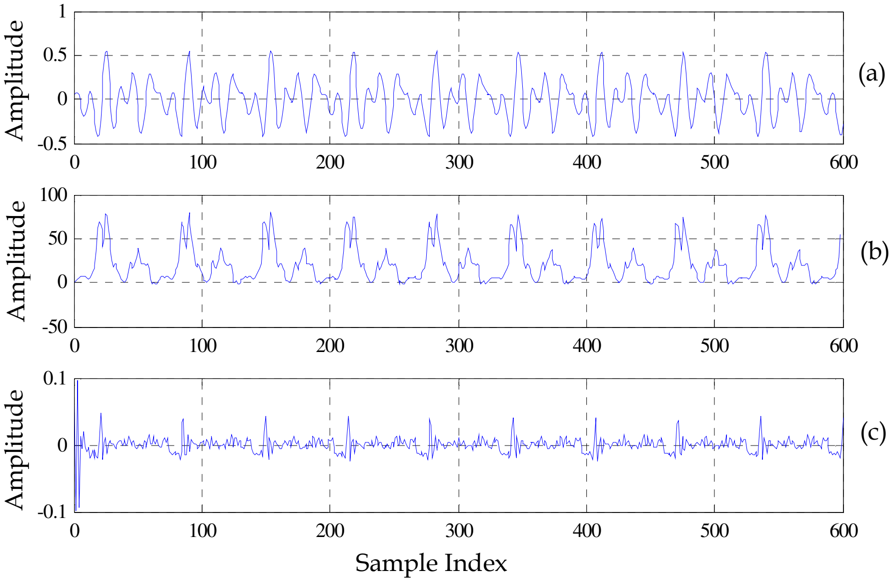


Figure 4.21. TEO analysis of voiced male speech signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a).

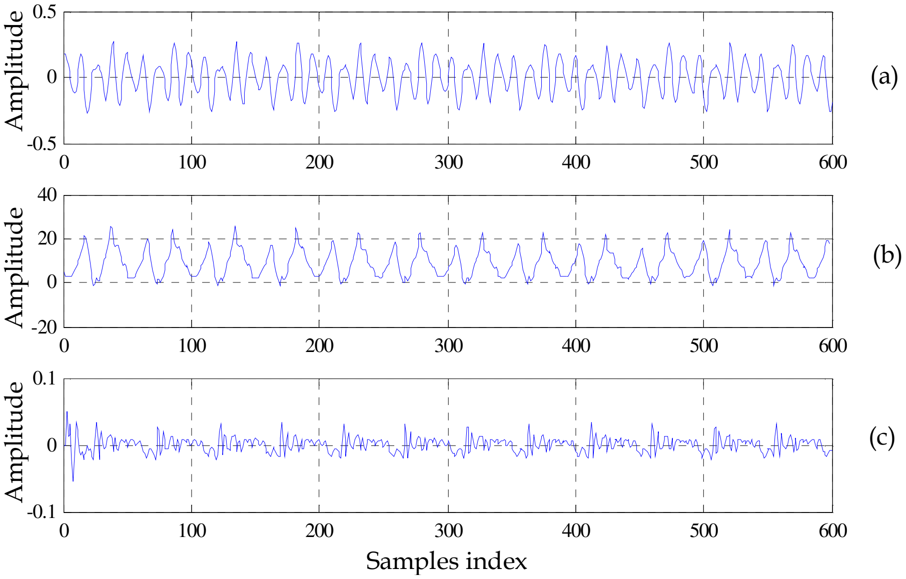


Figure 4.22. TEO analysis of voiced female speech signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a).

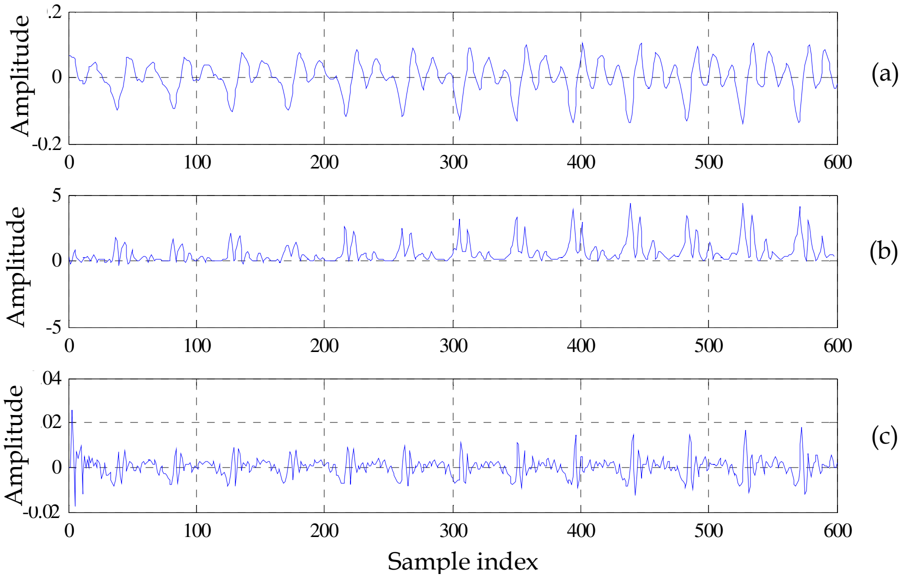


Figure 4.23. TEO analysis of voiced children speech signal, (a) time-domain signal (b) corresponding TEO profile and (c) LP residual of (a).

Following observations can be made from Figure 4.19 – Figure 4.23:

a. From Figure 4.19, it can be observed that the TEO as energy operator is not always positive, however, it is negative also at some places which depends upon whether $x^2(n) > x(n-1).x(n+1)$ or $x^2(n) < x(n-1).x(n+1)$.

b. During silence regions (as shown in Panel II Figure 4.19 (b)) TEO is zero, which indicates its ability to detect speech *vs.* silence regions also in the speech or infant cry segment. Thus, TEO has the capability to detect glottal activity (*i.e.*, vocal fold vibrations) *vs.* no glottal activity (*i.e.*, silence regions).

c. In all the speakers, it is observed that the maximum change in energy occurs at the glottal closure instant (GCI), which is due to the *sudden* closure of the vocal folds at GCIs (*i.e.*, impulse-like excitation which has large energy at GCI). This effect is the consequence of the sudden decrease in pressure at epochs. This is also observed from the LP residual of the signals. The peaks of LP residuals match with the peaks of TEO [**129**], indicating TEO profile also captures LP residual-like excitation source information. Thus, TEO profile of voiced speech signal can also be explored for F_0 or GCI estimation. In the next chapter (Section 5.7), a novel TEO-based method of F_0 extraction from infant cry signal is presented.

d. If the signal under consideration is a damped sinusoid then its TEO profile will be linearly decaying [**130**]. The presence of bumps in the TEO profile is an indicator of deviation from linearity. It can be observed from the TEO energy profiles of male, female, child speech and infant cry signals that the TEO energy profile is not smooth, it has many bumps within two consucutive GCI locations. This is an indicator of the non-linearity present in the speech production mechanism [**98**]. Hence, all speech or any sounds produced by a human being are a result of the nonlinear interactions of the vocal source and vocal tract.

e. It can be seen from the TEO profiles of the male, female, children and infant speakers (Figure 4.20 - Figure 4.23) that the variations in the TEO energy are different in all the speakers. It indicates different locations and manner of nonlinearity in speech production mechanism.

Thus, it can be said that though the vocal production system is not fully developed in infants, yet there is an indicator of the presence of airflow vortices via bumps in TEO profile within two consecutive GCIs which results in nonlinear behaviour of the voice production system in infants similar to that of adults.

4.7 Chapter Summary

In this chapter, effects of applying conventional and state-of-the-art signal processing methods on infant cry analysis are presented with suitable examples. It has been observed from the STFT analysis that it is difficult to identify formants and harmonics responses in the spectrum of the infant cry signal because of poor spectral resolution (due to the serious interaction of pitch harmonics with vocal tract spectrum). In the LP analysis of the infant cry signal, it has been observed that with the increase in LP order '*p*', LP spectrum peaks tends to match to the peaks of the pitch harmonics rather than formants which is primarily due to spectral matching mechanism while estimating optimum values of linear prediction coefficients (LPC). Along with this, the order of LP analysis is very low in case of infants compared to the adults because of the small vocal tract length. A similar effect is observed from the cepstral analysis of the infant cry signal where the lifter length is found to be very small to separate the vocal tract response from the infant cry signal. The infant cry signal analysis becomes more challenging because of the development of the vocal tract length in the early period of life of the infant. The LP order and the lifter size changes with the values of the formants and hence, depend on the length of the vocal tract which keeps on changing drastically in growing infants and children. Finally, we observed that TEO helps in analyzing nonlinearities associated with the production of infant cry signal. In addition, TEO profile is found to have the potential of detecting glottal activity (vocal fold vibrations) and no glottal activity in non-speech cry signals). In the next chapter, the spectrographic analysis is presented for different pathological infant cries, algorithms of fundamental frequency (F_0) estimation are applied on the infant cry signals and their analysis is shown and analysis of infant cries is attempted using prosodic features as well.

Chapter 5.

Analysis of Infant Cries

5.1 Introduction

The very first cry or the birth cry of an infant carries significant information about the health of an infant. As an infant grows, the acoustics of infant cry signal changes with the integration of vocal tract system and larynx. Infants are found to produce many sounds apart from crying, which reflect the learning mechanism of the infants of the language being spoken in his or her surroundings or the environment. Along with this, infants who have distinct cry sounds or who require a large amount of stimulation to produce a cry, are found to be at risk of sudden infant death syndrome (SIDS) or possible neurological disorders. In the initial portion of this chapter, different cry types (normal *vs.* pathological) are analyzed using the spectrographic analysis. For the spectrographic analysis, ten distinct cry modes defined by Xie *et. al.* are used [**40**]. Newborn infant cry is analyzed using features derived from fundamental frequency (F_0) or pitch contour, the energy of the cry signal in different frequency subbands and the relative amount of unvoicing present in the infant cry. For the extraction of the fundamental frequency, modified autocorrelation method is used and shown to perform better than traditional autocorrelation-based method. To identify the significance of these features in identifying the reason of crying, ANOVA analysis is applied to these features. It is observed that the F_0 features are not of much significance in the newborn cry analysis and presence of unvoicing in the infant cry varies with the maturity of central nervous system (CNS) and is a discriminative feature of prime importance in newborn's cry analysis. In birth cries, the mean percentage of unvoicing is *84.4* % which drops to *67.7* % in normal infants (*20* days -*3* months). Birth cry analysis shows that there is very less voicing and hence, less vibration of the vocal folds.

Features derived from the fundamental frequency (F_0) contour and energy contour are also used to analyze normal and pathological hunger and pain cries. Another method of F_0 extraction, namely, Teager Energy operator (TEO)-based method is proposed here in this analysis and is found to perform better than the autocorrelation-based method. In addition, the significance of unvoiced segments and fundamental frequency (F_0) in infant cry analysis is investigated. To find out the unvoiced segments from the infant cry, F_0 contour is used. For extraction of F_0 contour, TEO-based pitch extraction algorithm is used. TEO gives the running estimate of the signal's energy in terms of its amplitude and instantaneous frequency. To quantify the importance of proposed features in infant cry, analysis of variance (ANOVA) method is applied. Our results show that for cry classification, F_0-related features vary with the cause of cry, namely, hunger, pain, normal or pathological health condition of infants. These features can play an important role in the characterization of different cry types. Durational feature of cry is not found to be a significantly useful parameter for normal *vs.* pathological cry classification while it is a good feature for discriminating pain *vs.* hunger cries. It has been found that quantification of unvoiced segments and fundamental frequency (F_0) in the cry, deliver information about the maturation of cry production system. In infant cry analysis, the presence of high unvoicing ratio in a cry cannot be attributed to the presence of pathology (as may be the case with adult vocal fold pathological sounds).

5.2 Cryunit Segmentation

In this Section, cry recordings are divided into smaller cryunits for infant cry analysis. A cryunit is defined as the cry sound produced in one respiratory cycle [**55**]. The respiratory cycle has an inspiratory period and an expiratory period. Cry sound is produced during expiratory period only. In some cases, such as in the case of asthma, sounds are produced during the inspiratory period as well. However, the strength of the cry sounds produced during the expiratory period is higher than the energy of the sounds produced during inspiration. The spectrogram is used to identify the inspiratory and expiratory phases. Accordingly, cryunits are divided from the complete cry utterance as shown in Figure 5.1.

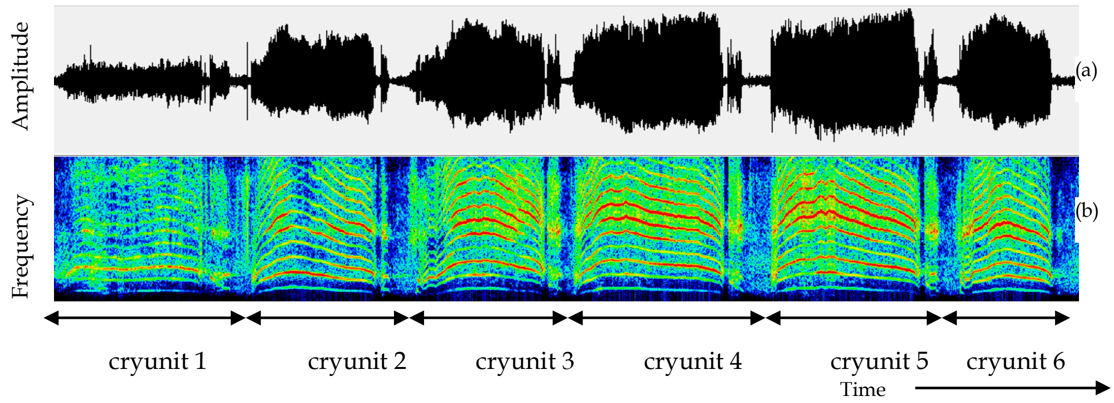


Figure 5.1. Formation of cryunits from a cry utterance (a) time-domain cry signal and (b) narrowband spectrogram of (a).

5.3 Spectrographic Analysis of Infant Cries

5.3.1 Introduction of Short-Time Fourier Transform (STFT) Analysis

The need for spectrographic analysis arises from the fact that Fourier analysis of a signal can be of use if the signal under study is stationary in nature [**38**], [**33**], [**131**]. However, the speech/ infant cry signal is a non-stationary signal and hence, cannot be studied effectively with Fourier transform. In STFT analysis, we divide the signal into smaller (comparatively) stationary segments using an analysis window and then Fourier analysis is performed on the smaller segment of the signal. The STFT analysis thus retains frequency as well as temporal information of the signal which represents a kind of *joint* time-frequency representation.

The spectrogram is the representation of the variation of signal energy along time and frequencies. Spectrograms are generally used in the fields of radar, sonar, music and speech processing. In the analysis of speech signals, the spectrogram is used for the identification of voiced, unvoiced and plosive sounds. Spectrograms are used to study the voice excitation source and vocal tract system.

To analyze the signal in frequency-domain Continuous Time Fourier transform (CTFT) is used. CTFT of a signal $s(t)$ is given by

$$S(\omega) = F\{s(t)\} = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt, \tag{5.1}$$

$$= \int_{-\infty}^{+\infty} s(t)\cos(\omega t)dt - j\int_{-\infty}^{+\infty} s(t)\sin(\omega t)dt. \tag{5.2}$$

The CTFT has infinite time-dimensional sine and cosine basis functions and hence, it shows poor resolution in time. Hence, instead of working with infinite-dimensional basis function, it is truncated to localize events in non-stationary signal or highly time-varying signal. This gives motivation to the introduction of short-time Fourier transform (STFT).

In *1946*, Dennis Gabor introduced the windowed Fourier transform to measure the frequency variations of the sound [**132**]. Spectrogram of a signal represents the squared magnitude of the STFT of a signal. Time is represented on the *X* - axis and frequency on the *Y*-axis. The spectrogram of signal $s(n)$ can be calculated as

$$S(n,\omega) = |\sum_{n=-\infty}^{\infty} s[n]w[n-m]e^{-j\omega n}|^2, \tag{5.3}$$

$$= |< s(n), w_{m,\omega}(n) >|^2, \tag{5.4}$$

where $s(n)$ is the sampled signal of $s(t)$, $w(n)$ is the analysis window, $w_{m,\omega}(n) = w[n-m]e^{j\omega n}$ and $< s(n), w_{m,\omega}(n) >$ indicates the inner product of signal $s(n)$ with time-frequency atoms, *i.e.*, $w_{m,\omega}(n) = w[n-m]e^{j\omega n}$.

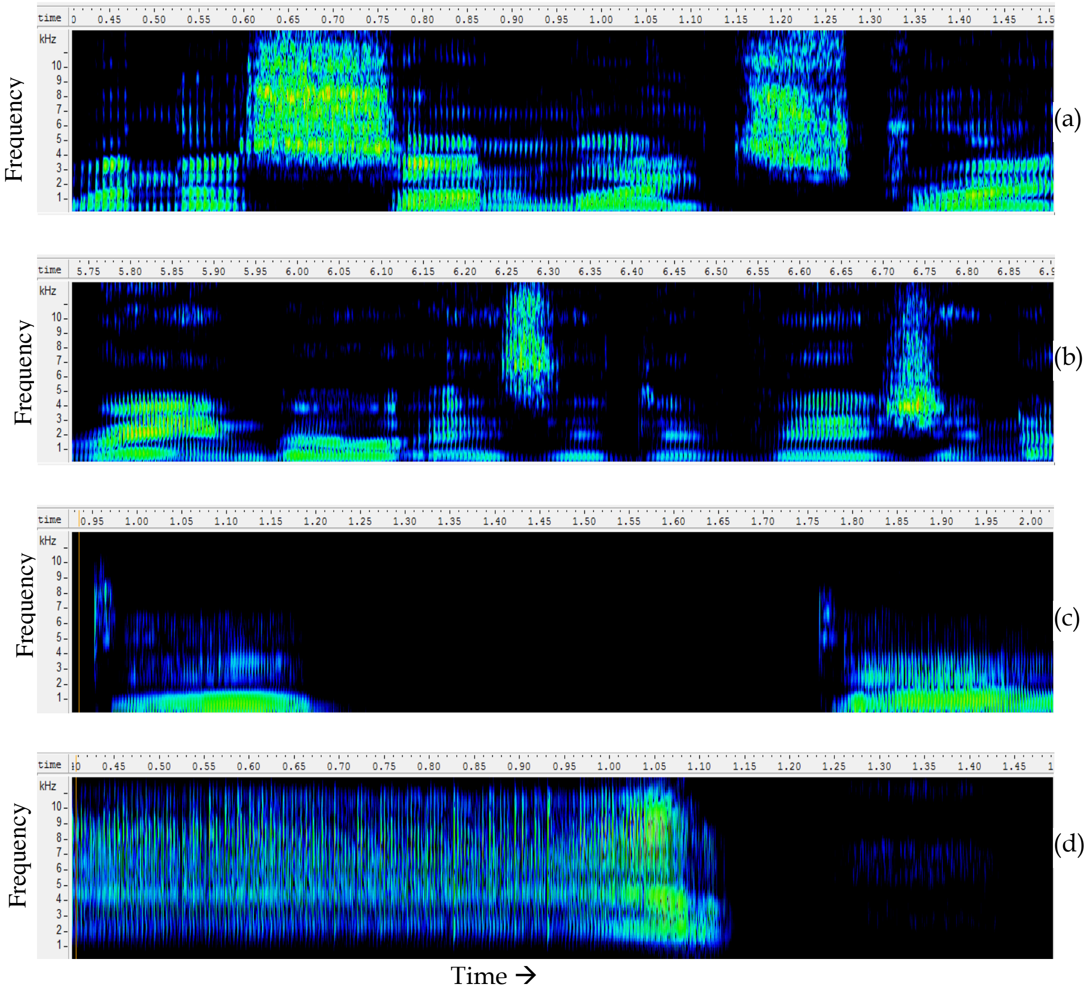


Figure 5.2. Wideband spectrogram of (a) normal male speech, (b) normal female speech, (c) normal child speech and (d) infant cry.

In the short-time Fourier transform (STFT), if the analysis window is taken less than a pitch period (< *3 ms*), then the resulting spectrogram is called the *wideband* spectrogram. If the window size is taken around *2-3* pitch periods longer (*i.e., 10-30 ms*), then the resulting spectrogram is called *narrowband* spectrogram. The narrowband spectrogram is used to define different types of vocalizations in birds, animals and human beings. An example of the narrowband and wideband spectrogram is shown in Figure 5.2. Narrowband spectrograms are used to investigate the properties of the excitation source (*i.e.*, harmonics associated with vocal fold vibrations) while wideband spectrograms show the characteristics of the vocal tract resonances, *i.e.*, formants and in particular, dynamic variations of formant contours (especially for F_1-F_4). The most common and straightforward analysis method that is used to calculate wideband spectrograms computes the spectrum over a short segment of the time-domain signal. Because of this, short-time segment (called a "window"), the analysis is able to capture rapid changes in the amplitude of the signal. For this reason, during voiced speech segments, the wideband spectrogram shows vertical lines corresponding to the rapid increase in amplitude that occurs when the vocal folds slap together. The narrowband frequency analysis is calculated over a much longer time window–too long to capture the rapid increase in amplitude that occurs at the time of vocal fold closure. Narrowband spectrograms have good frequency resolution. However, wideband spectrograms have a good temporal resolution.

Wideband spectrograms are generally used in speech signal processing-related applications such as word segmentation, phoneme segmentation, voicing, unvoicing and plosive detection. In Figure 5.3, wideband spectrograms of a male, female, child speech and infant cry of the same duration are shown for comparison.

STFT obeys the Heisenberg's uncertainty principle. The principle says that for a particle, more precisely the momentum is known, less precisely the position is known and vice-a-versa. The same principle can be applied in the signal processing framework for time-frequency representation of the signal. As has been stated earlier that the length of the window is directly proportional to the frequency resolution of the spectrum and it is inversely proportional to the temporal resolution of the signal in the time-frequency analysis. The uncertainty principle in the signal processing framework states that one cannot know what spectral components exists at what instance of time. However, one can know the time intervals in which a certain band of frequencies exists. This is known as time-frequency resolution [**133**]. Hence, if the signal length and window length are of same duration, we get good frequency resolution and temporal information is lost (*i.e.*, Fourier transform). Reducing the length of the window function improves the temporal information and reduces frequency resolution. Thus, there exist a trade-off between the window length and spectro-temporal resolution in STFT. For the spectrographic analysis, the selection of window size is also very critical. The uncertainty principle can be written as

$$\sigma_t^2 . \sigma_\omega^2 \geq \frac{1}{4}, \tag{5.5}$$

where σ_t^2 and σ_ω^2 are the spread in time and frequency-domain with zero mean, respectively. In particular, $\sigma_t^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} t^2 \,|\, f(t) \,|^2 \, dt$, $\sigma_\omega^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^2 \,|\, F(\omega) \,|^2 \, d\omega$ and $\| f(t) \|^2 = 1$.

In wideband spectrogram, vertical striations correspond to the local energy fluctuations. The rate of vocal fold vibrations is called fundamental frequency (F_0) of speech sound. It is known that the F_0 increases in the order defined as male, female, child and infant. From Figure 5.2 (a) as F_0 is low in male voice, the vertical striations are clear in the spectrogram. However, as we move from male to child voice, these are not at all clear and in infant's cry, these are very closely spaced and do not impart any significant information. From the wideband spectrograms of male and female voices, vocal tract resonances are clearly visible. However, in the case of child and infant cries, formants are *not* visible in the spectrogram. This is primarily due to the sampling of the vocal tract spectrum by the largely spaced pitch or excitation source harmonics. Thus, formant structure is there in the spectrum, however, we cannot see it in the computer due to signal processing artifacts of sampling. This is the reason that wideband spectrograms are not so much useful in infant cry analysis.

In Figure 5.3, narrowband and wideband spectrograms are shown for the same infant cry signal. In both the spectrograms, excitation source harmonics are dominating, making it difficult to identify vocal tract resonances in the wideband spectrogram which is primarily due to the serious interaction of pitch source harmonics with vocal tract spectrum, however, the excitation source and vocal tract harmonics are mixed together in the wideband spectrogram. On the other hand, in the narrowband spectrogram, the excitation source harmonics are clearly visible and hence, these can be used to define various infant cry modes (*e.g.*, the study reported in [**38**]). Therefore, in the remaining portion of this chapter, narrowband spectrograms are used for infant cry analysis.

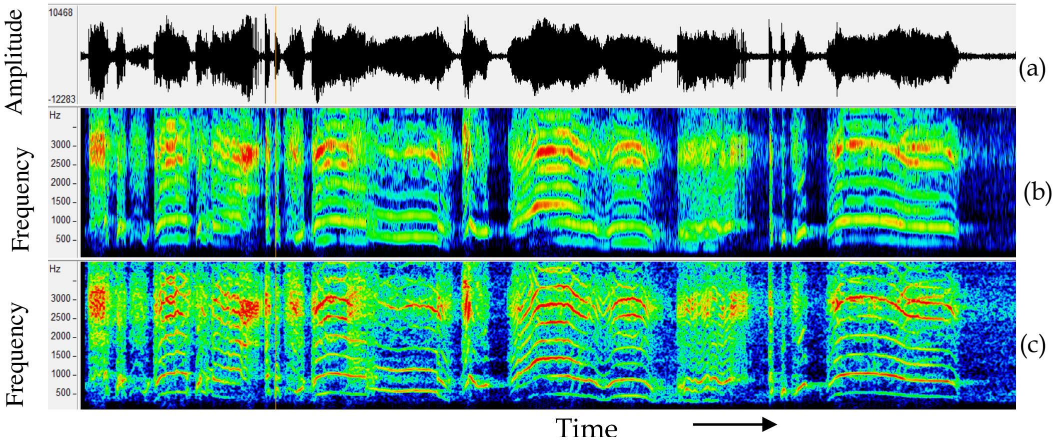


Figure 5.3. (a) Time-domain waveform of the cry signal (b) wideband spectrogram and (c) narrowband spectrogram.

5.3.2 Selection of Window Length

The resolution of the spectrogram depends on the time and frequency of the window used in the STFT analysis. Specifically, it depends on the area of the Heisenberg's box, *i.e.*, $\sigma_t^2.\sigma_\omega^2$. As mentioned above, the length of the window is directly proportional to the frequency resolution and inversely proportional to the temporal resolution. The same has been observed in our experiments conducted on infant cry database for varying window durations. It can be observed from the Figure 5.4 that using a window length of *1.5 ms* gives a wideband spectrogram while changing the window length to *5 ms* gives a better representation of formant contour in the spectrogram. Furthermore, increase in window length from *5 ms* to *20 ms* improves the energy distribution among excitation source harmonics and clear harmonic structure is visible in the produced narrowband spectrogram.

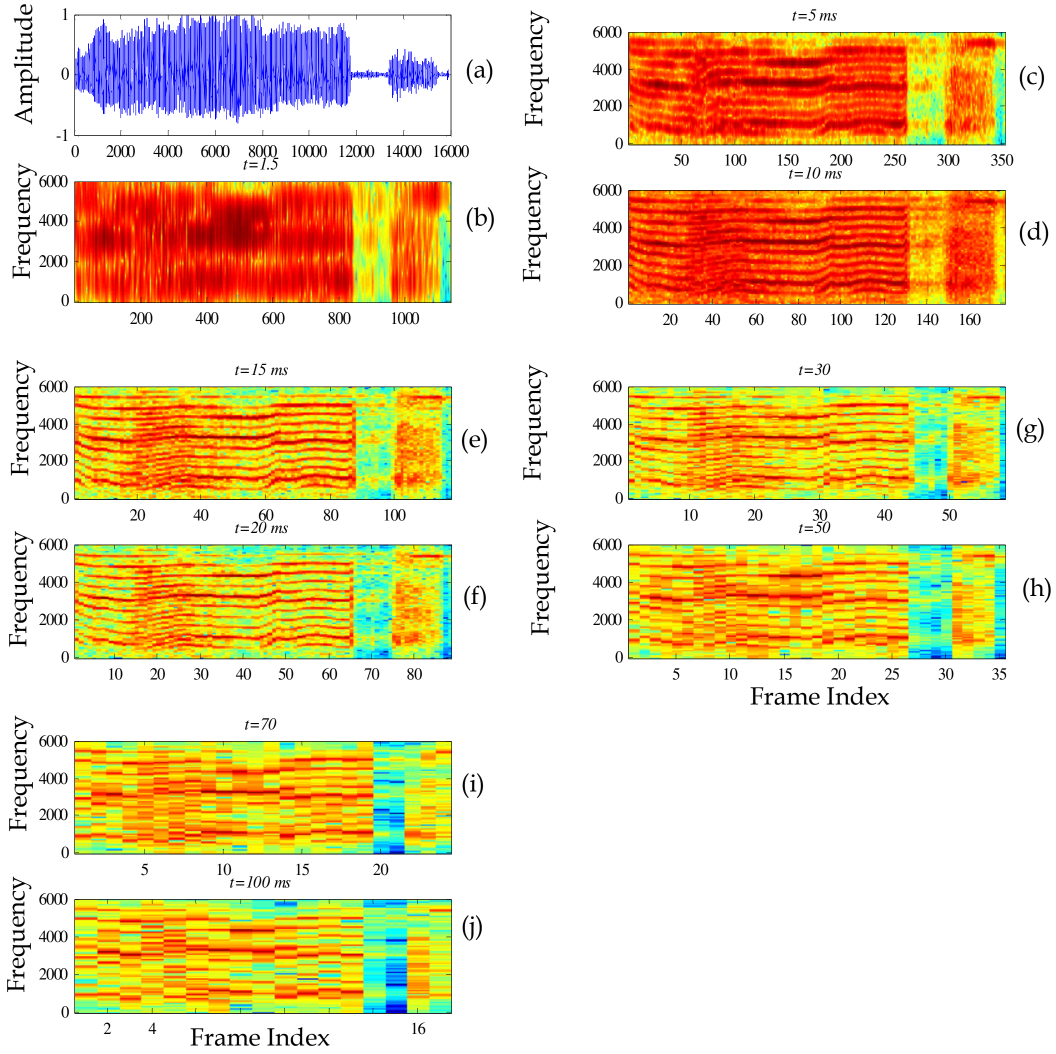


Figure 5.4 Effect of variation in window length on the STFT analysis. (a) time-domain infant cry signal and its narrowband spectrogram with window size of (b) *1.5 ms*, (c) *5 ms*, (d) *10 ms*, (e) *15 ms*, (f) *20 ms*, (g) *30 ms*, (h) *50 ms*, (i) *70 ms* (j) *100 ms*.

We can observe that increasing the window length beyond *20 ms* makes the time-frequency resolution very poor and it becomes difficult to extract any useful information from these spectrograms. In all these spectrograms, Hamming window function is used and to find out the reason of poor resolution, the frequency characteristics of the Hamming window for varying window length are plotted in Figure 5.5.

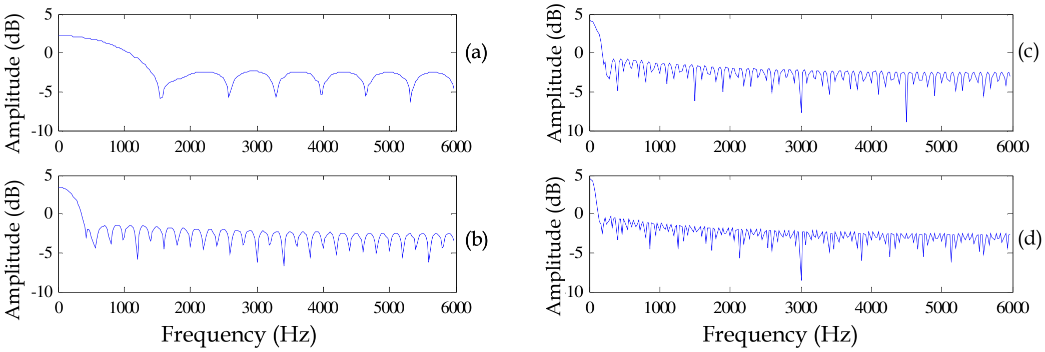


Figure 5.5 Effect of variation in window length on the frequency response characteristics of the window. Window lengths are (a) *1.5* ms, (b) *5* ms (c) *10* ms and (d) *20* ms.

It is observed from the figure, as the window length is small (say *1.5 ms*), the number of sidelobes is small and the sidelobes are located around the formant locations. Hence, information about the energy distribution of cry around formant locations appears in wideband spectrograms. Increasing the window length, increases the number of sidelobes and reduces the width of the main lobe, thereby, results in a harmonic structure in the spectrogram. Further increase in the window length results in dominating sidelobes and is reflected as very blurred image in the spectrogram

5.3.3 Selection of Window Function

To observe the effect of the various window functions on the spectral resolution of the spectrogram, experiments have been done and results are reported in Figure 5.6. The window functions used in this study are reported in Table 5.1. In Table 5.1, $\Delta\omega$ is the mean-square bandwidth and *A* represents the maximum amplitude of the first sidelobe located at $\omega = \omega_0$ (where ω_0 is the frequency where mainlobe has the minimum value). In all the spectrograms, the window length is kept constant at *20 ms* and an overlap of *0.75* times the window length duration. From Figure 5.6, it can be observed that the selection of window in the spectrographic analysis is also an important parameter. For example, in the case of the rectangular window, because of poor separation of main lobe and sidelobes, the frequency resolution of the spectrogram is poor and it appears smearing. However, Hamming window and Gaussian window produces a very clear harmonic structure in the narrowband spectrograms as these windows have clear and wide mainlobe which is separated from the sidelobes by large amplitudes. Though, flattop window has widest mainlobe and very small sidelobes (as shown in Figure 5.7), it does not seem to be good in the analysis of infant cries.

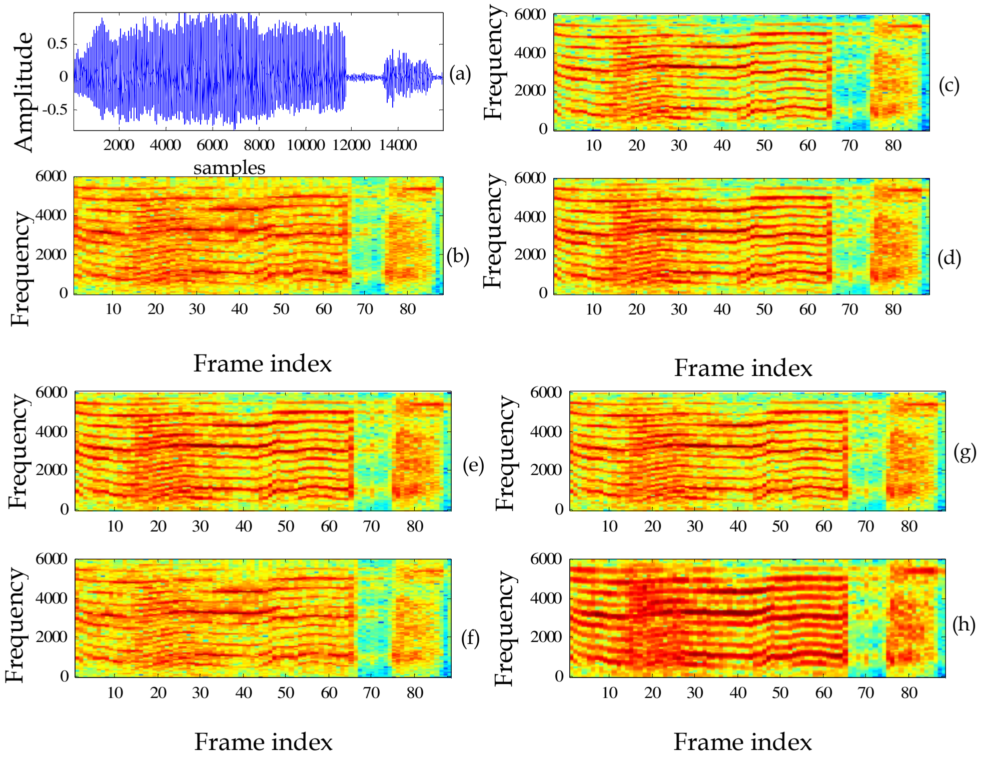


Figure 5.6 Effect of selection of window on the spectrogram. (a) Time-domain signal and its corresponding spectrogram using (b) rectangular (c) Hamming (d) Hanning (e) Gaussian (f) Kaiser (g) triangular (h) flat top windows.

Table 5.1. Window functions used in the analysis

Window function	Window function	$\Delta\omega$	A (dB)
Rect-angular	1	0.89	-13
Hamming	$0.54-0.46\cos(2\pi t)$	1.36	-43
Hanning	$\cos^2(\pi t)$	1.44	-32
Gaussian	$\exp(-18t^2)$	1.55	-55
Kaiser	$I_0(\pi\alpha(\sqrt{\frac{1-(\frac{2n}{n-1}-1)^2}{I_0(\pi\alpha)}})$	-	-
Triangular	$1-\lvert\frac{n-\frac{N-1}{2}}{\frac{L}{2}}\rvert$	-	-
Flat top	$a_0-a_1\cos(\frac{2\pi n}{N-1})+a_2\cos(\frac{4\pi n}{N-1})-a_3\cos(\frac{6\pi n}{N-1})+a_4\cos(\frac{8\pi n}{N-1})$, $a_0=1$, $a_1=1.93$, $a_2=1.29$, $a_3=0.388$, $a_3=0.028$	-	-

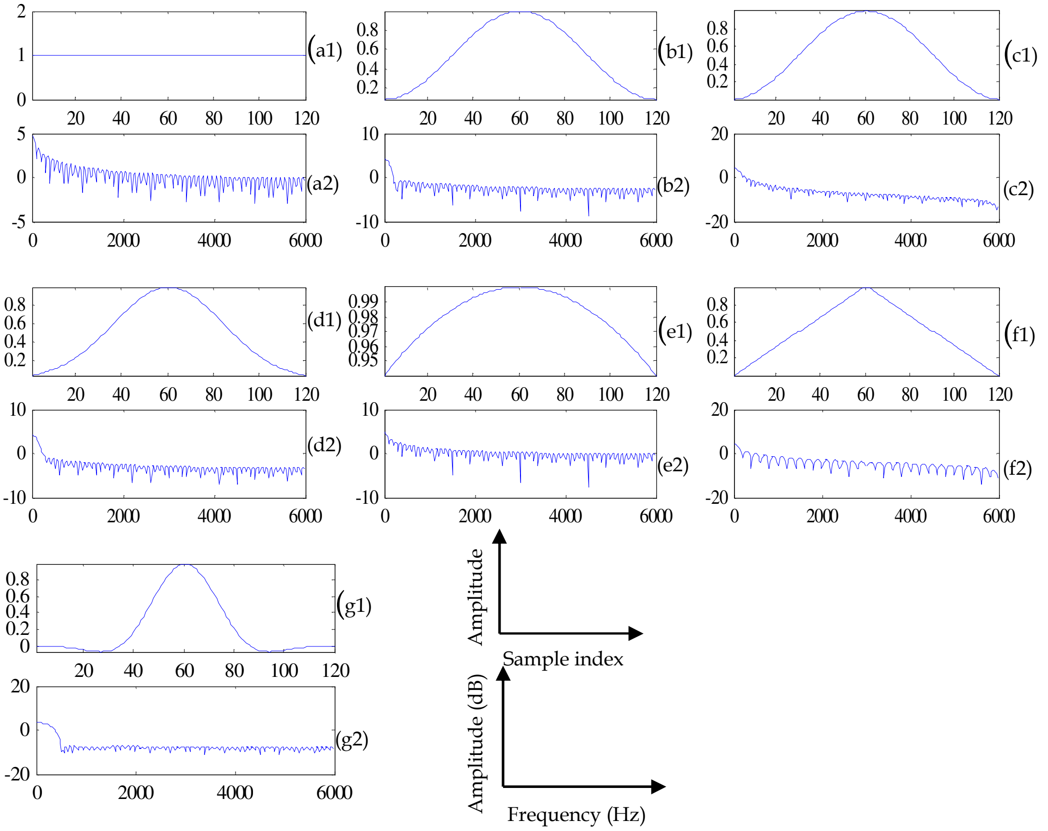


Figure 5.7. Various window functions and their frequency responses. (a) Rectangular window (b) Hamming (c) Hanning (d) Gaussian (e) Kaiser (f) triangular and (g) flattop windows.

We can observe from the spectrogram where flattop window is used, that it gives excellent harmonic structure compared to the Hamming and Gaussian windows. However, it is difficult to identify double harmonic break regions (frame index *20-30*) with this window because of very wide spectral main lobe there is interference in the harmonics present in the double harmonic breaks, and it appears like noise in the spectrogram. Thus, after all these considerations, Hamming window is found to be the best choice for the infant cry analysis work and hence, used for all spectrographic analysis in the remaining part of this chapter.

5.3.4 Infant Cry Modes from Narrowband Spectrogram

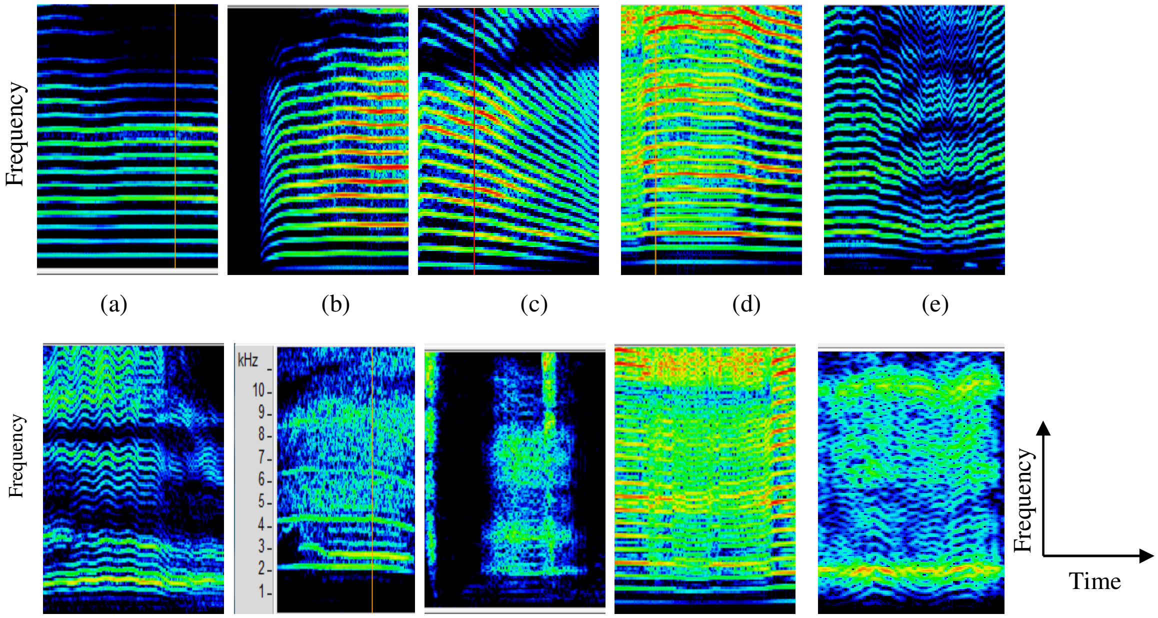
From the spectrographic patterns of the infant cries, various cry modes have been identified by many researchers. The ten distinct cry modes used here for the analysis of infant cries are as follows [**40**], [**94**]:

a. **Flat**: Flat melody pattern is characterized by approximately constant F_0 with time. The harmonics in this region are clearly observable.
b. **Rising**: This is characterized by an increase in F_0 with time. The harmonics are clearly visible in this region.
c. **Falling**: Similar to rising melody pattern except that the F_0 decreases with time.
d. **Double harmonic breaks**: Simultaneous parallel lines are observed in between the harmonics of the F_0. These harmonics are of a frequency other than the F_0 of the vocal excitation source. The presence of double harmonics breaks is generally correlated with the pathological infant cries. However, these are also observed in newborn infant cries.
e. **Glottal roll or glide**: It occurs at the end of the expiratory phonation. This shows the vibratory pattern of the F_0 and its harmonics. The energy in these harmonics decreases gradually.

f. **Weak vibration or vibrato**: It is similar to the glottal roll except that this may occur in the middle of the infant cry also instead of at the end of the cry. The energy in weak vibrations is much smaller.

g. **Hyperphonation**: It is defined as the regions where F_0 exceeds 1 kHz. The presence of hyperphonation is related with the presence of pathology (*i.e.,* neural disorders).

h. **Inspiratory phonation**: It occurs due to the sound made by the infant during inhalation. This occurs before the phonation region of the cry sound. Typically of much smaller duration.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 5.8. Infant cry modes from narrowband spectrogram (a) flat (b) rising (c) falling (d) double harmonic break (e) glottal roll (f) vibrations (g) hyperphonation (h) inspiratory phonation (i) dysphonation and (j) weak vibration.

i. **Dysphonation**: This is the noise concentration found in the infant cries. This is characterized by the irregular or unstructured distribution of energy and typically the energy in this region is very high and heavy *turbulence* is created in this region. If dysphonation dominates in the spectrogram, it may indicate the presence of pathology. However, newborn infant cries also have high dysphonation regions.

j. **Vibrations**: These are similar to weak vibrations, however, occur with high energy.

All these cry modes are shown in Figure 5.8. Next, spectrographic studies were carried out for infants suffering from laryngomalacia, asphyxia, deafness, meningitis, brain hemorrhage, *etc*.

5.3.5 Spectrographic Analysis and Observations

5.3.5.1 Normal infant cry

Doctor's comment during recording of infant cry: The normal infant's cry was recorded during vaccination. The infant cry is a pain cry of a 3 month old infant.

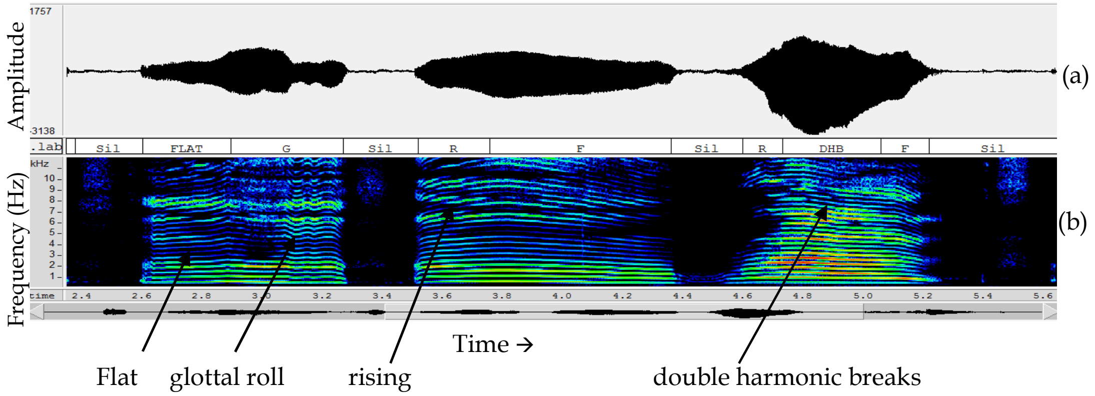


Figure 5.9. (a) Time-domain waveform, (b) corresponding spectrogram and cry modes present in the spectrogram of a normal infant's cry.

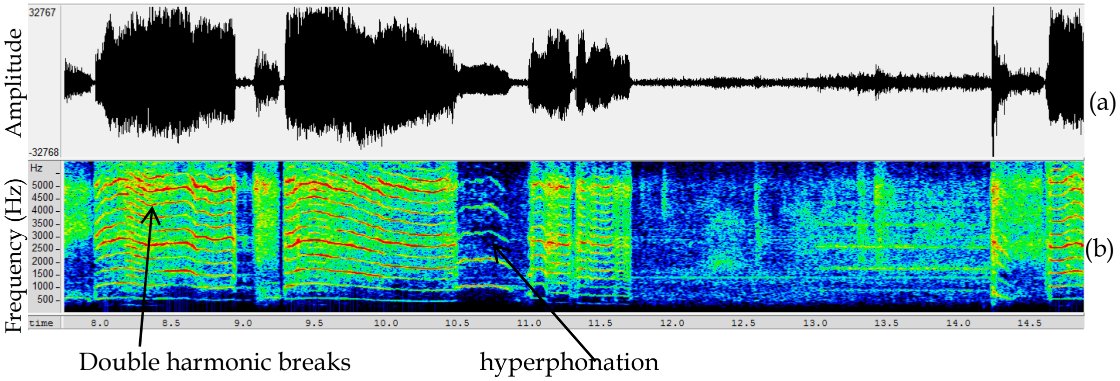


Figure 5.10. (a) Time-domain waveform, (b) corresponding spectrogram and cry modes present in the spectrogram of a normal infant's cry.

Spectrographic Analysis: Narrowband spectrograms of normal infant cry are shown in Figure 5.9 and Figure 5.10. In normal infant‘s cry, melody pattern is flat or rising/falling (*i.e.*, rising in the beginning and falling at the end of expiratory phase) and glide is present. Hyperphonation and double harmonic breaks are also seen. Pitch harmonics are clearly visible and constant.

5.3.5.2 *Neonatal cry*

Doctor‘s comments during recording of the cry: This is the cry of a normal infant. The infant is 5 days old and it is a pain cry due to injection.

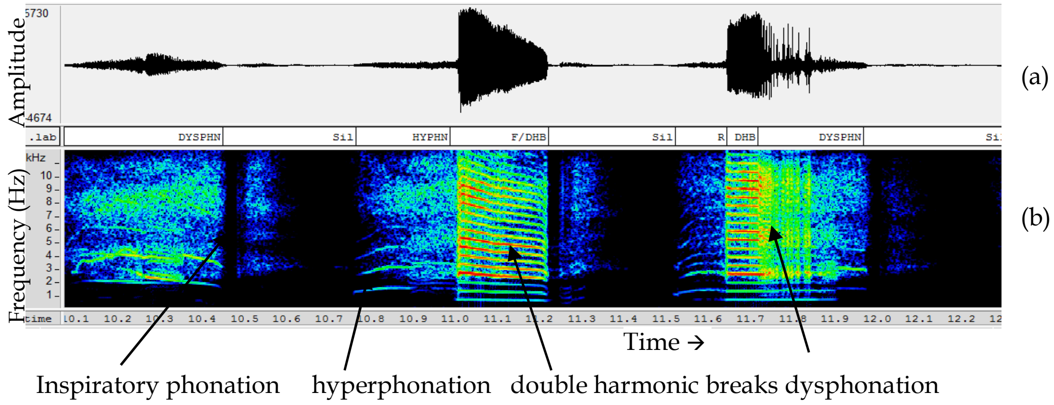


Figure 5.11. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of a neonate’s cry.

Spectrographic Analysis: It has a very small duration of cryunits and very small energy because of poor control of human brain on rib cage movement. Spectral characteristics are shown in Figure 5.11 and same as normal cry with the presence of *double harmonic breaks* and very frequent inhalation pattern. Hyperphonation is also seen in the cry.

5.3.5.3 *Cry in infants with Larynx not Developed (Laryngomalacia)*

Symptoms and Cause: It is a common cause of congenital stridor and a kind of abnormality of laryngeal cartilage. An infant with this abnormality produces typical sound (noisy breathing), which is mostly unvoiced. It may represent a delay of maturation of the supporting structure of the larynx [**134**].

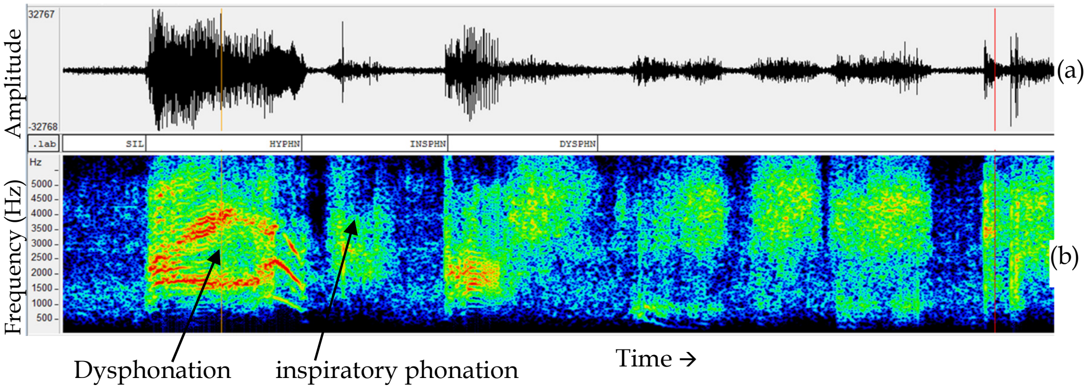


Figure 5.12. (a) Time-domain infant cry waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from laryngomalacia.

Spectrographic analysis: From spectrogram shown in Figure 5.12, it is observed that

1. Dysphonation, hyperphonation and inhalation are dominating in this cry.
2. Double harmonic break, glottal roll and glide are totally absent.
3. Spectral resolution is poor due to severe turbulence (and thus, unstructured energy distribution) in cry signal.

Similar observations were made in [**135**].

5.3.5.4 Infant with Asthma

Doctor's comment during recording of the cry: This is a cry of an asthematic patient. Now comes with accute bronchial asthma. This is an abnormal cry.

Symptoms and Causes: Asthma is a chronic inflammatory disease of the airways in which the airways become blocked or narrowed. An infant suffering from asthma has shortness of breath and whistling sound while crying [**136**].

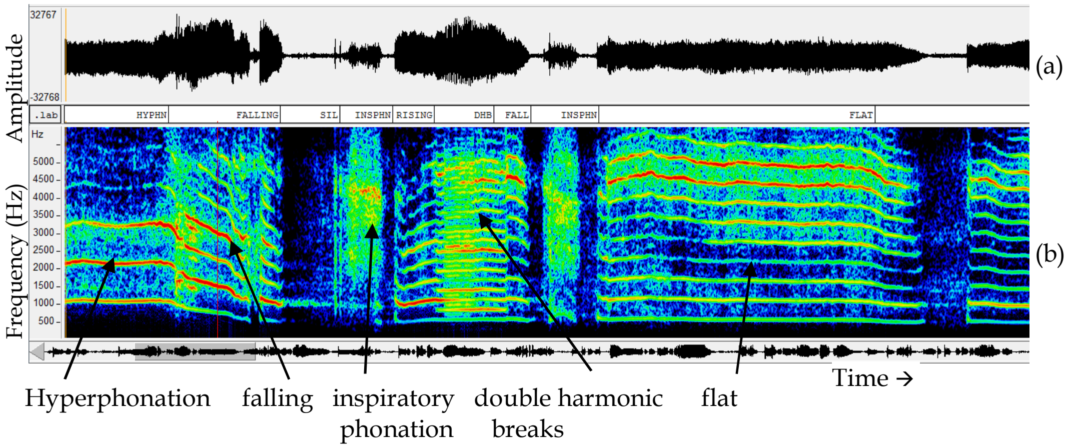


Figure 5.13. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from asthma

Spectrographic Analysis: Due to frequent inhalation, inspiratory phonation is observed in the spectrogram shown in Figure 5.13. Other cry modes are mostly same as a normal cry. Double harmonic breaks are visible in the spectrogram of the cry.

1. Because of the problem in breathing, inhalation is frequent in the spectrogram.
2. Rising, falling and hyperphonation cry modes are present. It is similar to normal infant cry.

5.3.5.5 *Congenital heart disease*

Doctor's comment during recording of the infant cry: This infant is suffering from congenital heart disease. The infant is 4 months old boy and is crying because of hunger.

Symptoms and causes: Congenital heart defect is a birth defect of heart structure. It is the most common type of birth defect. This defect may be in the walls of heart, arteries, veins or with the direction of blood flow. The infant may or may not show symptoms in early infancy [**137**].

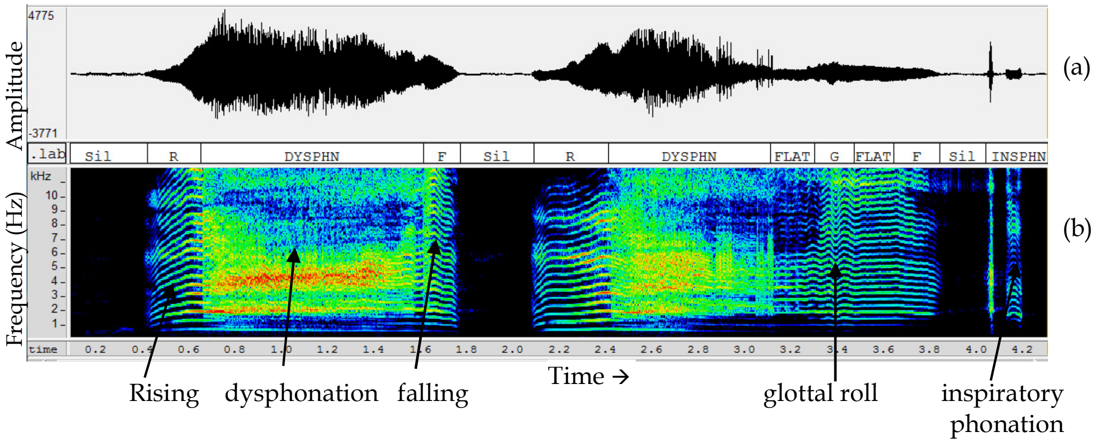


Figure 5.14. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from congenital heart disease.

Spectrographic Analysis: Spectrogram of an infant cry suffering from congenital heart defect is shown in Figure 5.14. It can be observed that the melody type is rising followed by falling, same as normal infant. Glide is present and dysphonation is dominating in the spectrogram.

5.3.5.6 Infant with Down syndrome

Doctor's comment during recording of the cry: This infant is suffering from Down Syndrome which is a disease caused by *chromosomal* abnormality. He is crying because of hunger. He is *9* months old boy child.

Symptoms and causes: It is also known as trisomy *21*. It is a genetic disorder caused by chromosomal abnormality. Infant suffering from this disorder shows poor developmental and intellectual growth [**138**].

Spectrographic Analysis: Spectrographic analysis shown in Figure 5.15, shows that such infant cry has flat melody pattern with no glide. The duration of the cries of such infants is longer than normal infants. During inspiration, silence is observed.

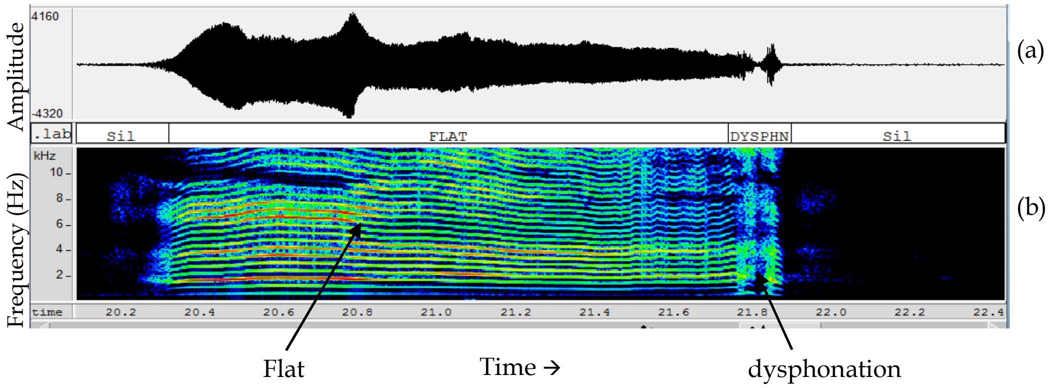


Figure 5.15. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from Down's syndrome.

5.3.5.7 *Brain hemorrhage*

Doctor's comment during recording of the cry: The infant is diagnosed with subdural hemorrhage, in which some part of the brain tissues are permanently damaged. The boy child is one year old and he is crying because of hunger.

Symptoms and causes: Intraventricular hemorrhage of the newborn occurs due to bleeding in the fluid filled areas inside the brain [**139**].

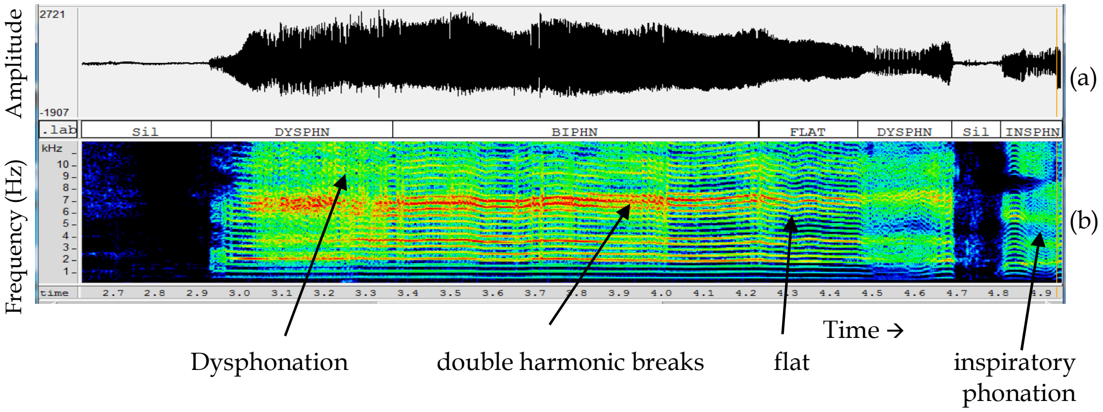


Figure 5.16. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and present in the narrowband spectrogram of an infant suffering from brain hemorrhage.

Spectrographic Analysis: From Figure 5.16, it is observed that the melody pattern of cry is flat. Noise concentration (dysphonation) is higher and double harmonic breaks are present. Sometimes inspiratory phonation is also present in the cry.

5.3.5.8 *Malnutrition*

Doctor's comment during recording of the infant cry: The infant is diagnosed with malnutrition. The boy child is *20* months old and he is crying because of hunger.

Symptoms and causes: Malnutrition is the condition caused by poor nutrient supply in the food. Effected infant is at higher risk of infection and infectious diseases. Malnutrition weakens the immune system. The risk for death increases with increasing level of malnutrition. Infants suffering from it show poor or over weight gain and poor physical development [**140**].

Spectrographic analysis: Spectrogram of an infant cry with malnutrition is shown in Figure 5.17. It is observed that melody pattern is mostly same as a normal infant. Glide is rare in these cries. The energy concentration in the higher frequency region is very poor. Sometimes inspiratory phonation is also present. Double harmonic breaks are not seen in this type of infant cry. Research on these infants shows that changes in the cry characteristics depends on the degree by which brain gets effected due to malnutrition [**141**].

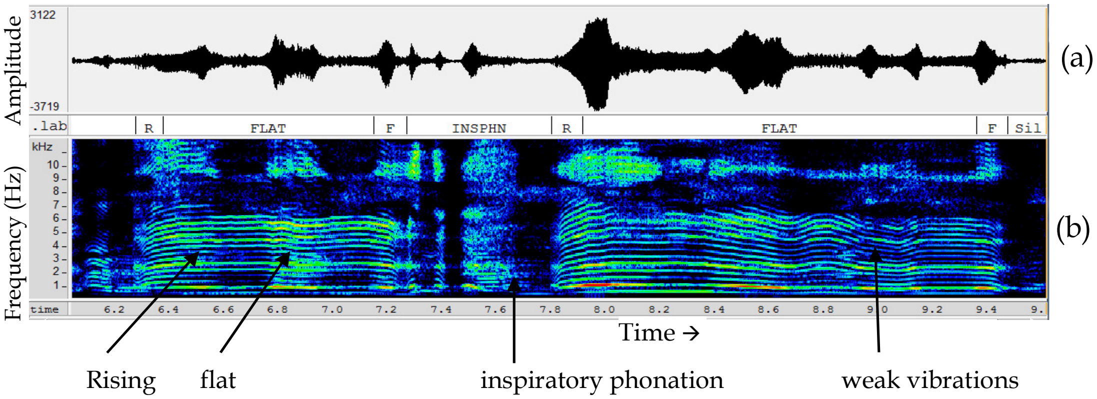


Figure 5.17. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from malnutrition.

5.3.5.9 *Hypoxy Ischemic Encephalopathy (HIE)*

Doctor's comment during recording: This is a newborn's cry (*2* days old newborn) who is suffering from HIE with seizures.

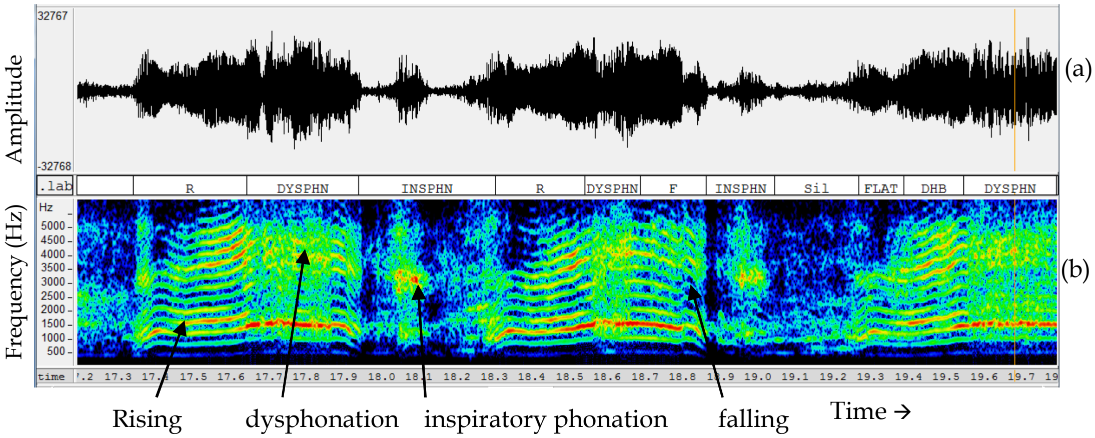


Figure 5.18. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from HIE.

Symptoms and causes: Brain damage from severe oxygen deficiency after the birth. If the oxygen deficiency in the brain cells continues, the brain tissues destroys and this may result in motor and mental handicap of the child [**142**].

Spectrographic Analysis: Melody pattern of such cries are rising followed by falling. Inspiratory phonation and dysphonation are seen in the spectrogram of the cry (as shown in Figure 5.18). Duration of cry is less.

5.3.5.10 Hydrocephalus

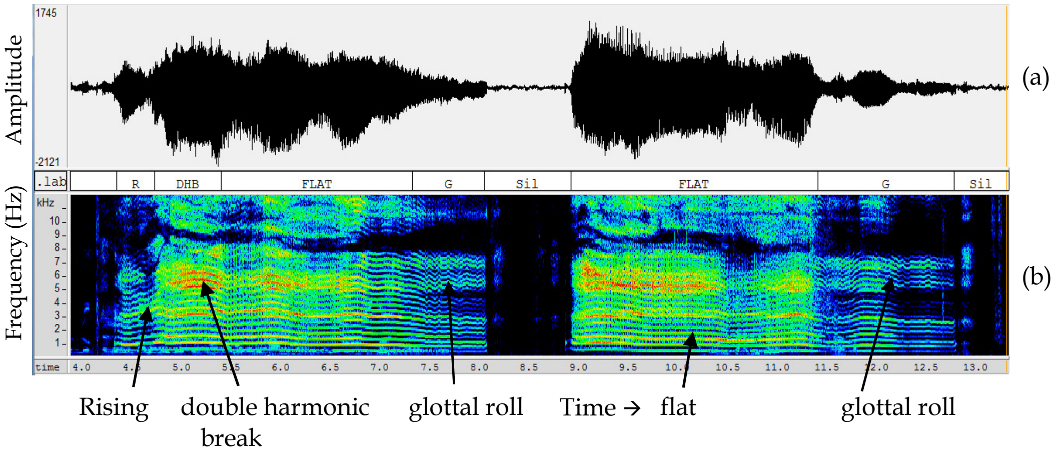


Figure 5.19. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from hydrocephalus.

Doctor's comment during recording of the cry: This is the cry of an *11* month old infant. This boy is suffering from hydrocephalus, a condition in which a fluid is accumulated in the human brain. He is crying because of discomfort.

Symptoms and causes: Hydrocephalus is the condition in which there is an excessive accumulation of the cerebrospinal fluid in the brain. This excessive fluid causes abnormal widening of the spaces in the brain. This widening creates excessive pressure on the human brain cell tissues [**143**].

Spectrographic Analysis: It can be observed that melody pattern is mostly flat. Glottal roll and double harmonic breaks are present. Inspiratory phonation is not observed in the infant cry. In the high frequency band, energy concentration is very low. Duration of infant cry is also observed to be longer than normal infant's cry as shown in Figure 5.19.

5.3.5.11 Meningitis

Doctor's comment during recording of cry: Infant is suffering from pyomeningitis. Pyomeningitis is a case of remote symptomatic epilepsy due to mismanaged feeding of external milk and feeding with the bottle. The infant is a boy child of *6* months age and he is crying because of hunger.

Symptoms and causes: Bacterial meningitis is an inflammation of the membrane that covers the brain and spinal cord, called *meninges*. Bacterial meningitis can be life threatening. The symptoms are fever, lethargy, stiff neck, rashes on skin and seizures [**144**].

Spectrographic Analysis: From Figure 5.20 (b1), it can be observed that glottal roll and dysphonation are dominating in the spectrogram. During the silence, inspiratory phonation is absent. Duration of cry is very short and energy is also very low. For the same pathology, another infant's cry is being analyzed using spectrogram (as shown in Figure 5.20 (b2)). The duration of cry is longer than normal infant cry. Melody pattern is flat in both the spectrograms with the presence of dysphonation and inspiratory phonation.

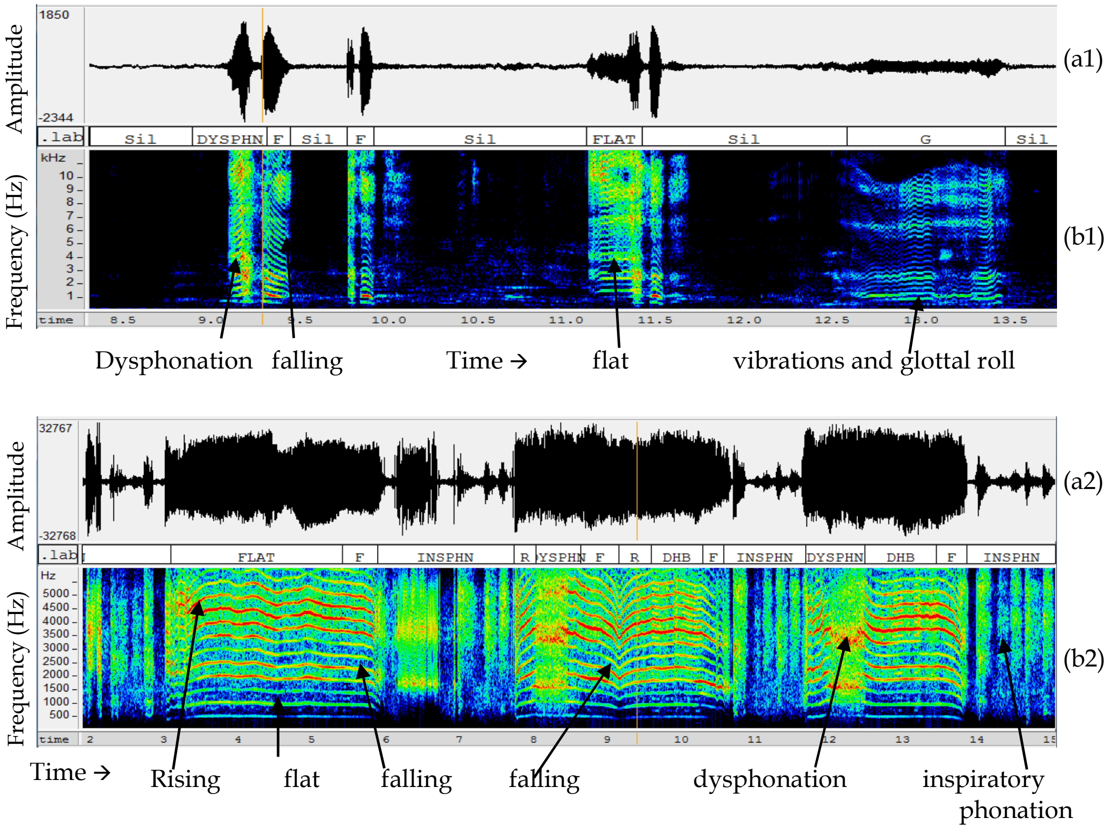


Figure 5.20. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of two different infants suffering from meningitis.

5.3.5.12 Respiratory distress syndrome (RDS)

Doctor's comment during recording of cry: This is the cry of an infant suffering from respiratory distress. This is an abnormal cry.

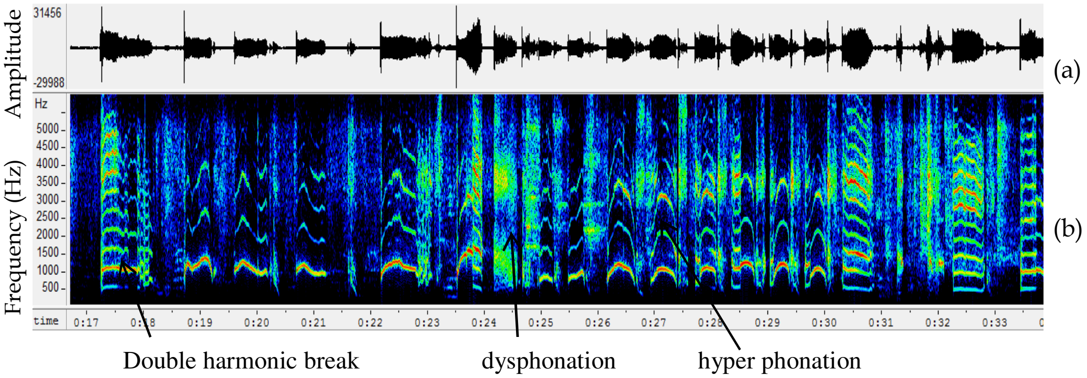


Figure 5.21. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from respiratory distress.

Symptoms and causes: Respiratory distress syndrome often occurs in the premature babies (< *40* weeks of GA). This occurs due to underdeveloped lungs in the babies. The symptoms are rapid breathing, apnea, shortness of breath and grunting sounds while breathing [**145**].

Spectrographic Analysis: Analysis of spectrogram of a cry of an infant suffering from respiratory distress shows following observations from Figure 5.21:

1. Duration of cry is short.
2. Inspiratory phonation is present due to the problem in breathing.
3. Double harmonic breaks are present and
4. Concentration of energy is less.

5.3.5.13 Deaf infant's cry

This sample is taken from the *Corpus III* (discussed in Chapter 3).

Symptoms and causes: The major cause of hearing loss is an infection. Other important reasons are genetics and development of diabetes in the mother during pregnancy. Newborns with hearing loss do not startle by a loud sound, an older infant does not show reactions to different sounds and sound levels [**146**].

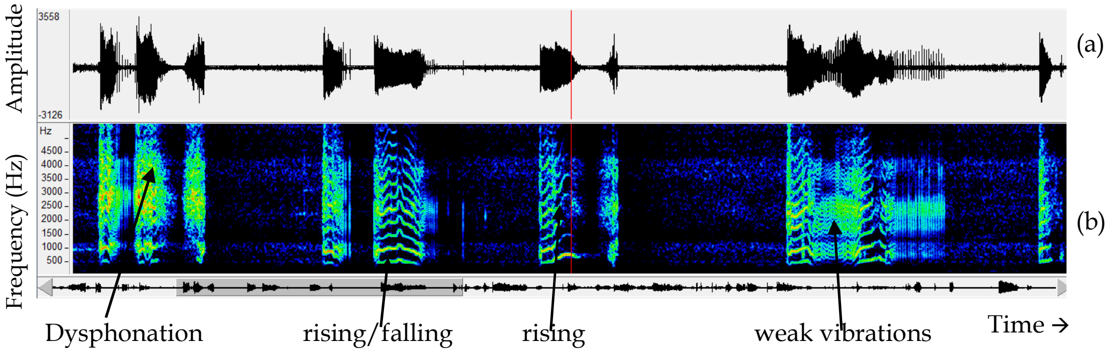


Figure 5.22. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of a deaf infant's cry.

Spectrographic Analysis: In the spectrogram of the deaf infant cry shown in Figure 5.22, we can observe that the cry is of very short duration followed by a long silence interval. Source harmonics are visible only in some part of the infant cry. Dysphonation is dominating the spectrogram. Inspiratory phonation is not observed in the cry. Generally, the melody (prosodic) pattern is rising followed by sharp fall. Weak vibrations are also observed in the spectrogram.

5.3.5.14 *Asphyxia*

The infant cry sample is taken from the *Corpus III.*

Symptoms and causes: It is a condition caused by insufficient oxygen after the birth. It is the inability to breath normally and may cause damage to brain tissues due to oxygen deficiency. Infants suffering from asphyxia show poor heart rate, pale skin, poor muscle tone and infant may suffer from seizures [**147**].

Spectrographic analysis: The spectrogram show very less amount of energy in the infant cry. Harmonics are clearly visible in the spectrogram. Vibrations and weak vibrations are observed in the infant cry as shown in Figure 5.23.

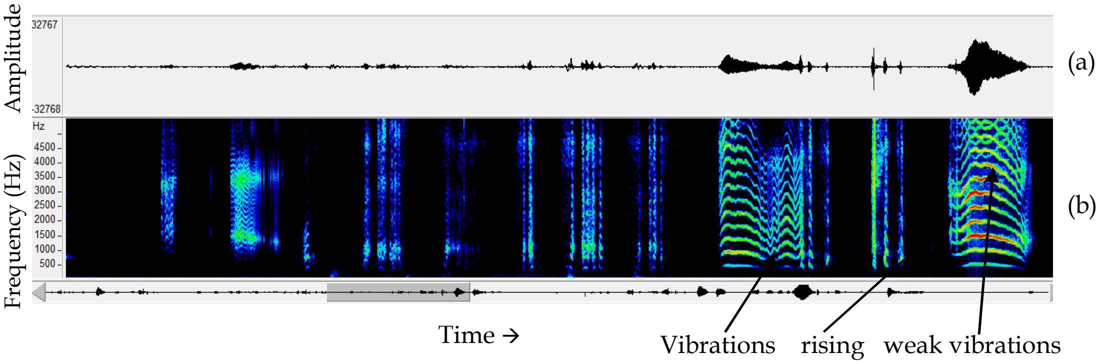


Figure 5.23. (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from asphyxia.

5.3.6 Summary of Observations from Spectrographic Analysis

A summary of presence or absence of these modes is given in Table 5.2.

Table 5.2. Presence of different modes in the spectrogram of a cry

Cry Modes Pathology	Flat	Rising	Falling	Double harmonic breaks	Glottal roll	Weak vibration	Hyper phonation	Inspiratory phonation	Dysphonation	vibrations
Normal	P	P	P	P	P	N	N	N	N	P
Neonatal	P	P	P	P	N	N	N	P	P	P
Laryngomalacia	N	P	P	N	N	N	P	P	P	N
Asthma	P	P	P	P	N	N	N	P	N	N
Heart disease	N	P	P	N	P	N	N	N	P	N
Down syndrome	P	N	N	N	N	N	N	N	N	N
Malnutrition	P	P	P	N	N	N	N	P	N	N
HIE	P	P	P	P	N	N	N	P	N	N
Hydrocephalus	P	P	P	P	P	N	N	N	N	N
Meningitis	P	P	P	N	P	N	N	P	P	P
RDS	P	P	N	P	N	N	P	P	N	N
Deaf	N	P	P	N	N	P	N	N	P	N
Asphyxia	N	P	P	N	N	P	N	N	N	N
Brain hemorrhage	P	N	N	P	N	N	N	P	P	N

P=Present, N= Absent

From the spectrographic analysis of different cries of several pathologies presented in this Section, it can be concluded that with the presence and severity of the pathological condition, dysphonation relatively dominates in the infant cry. The duration of the cry may become longer or smaller. Inspiratory phonation and hyperphonation are present mostly in the pathological cases (sometimes, it is observed in newborn cries also); these modes are not observed in the normal infant's cry. It can be observed from Table 5.2 that in the case of pathological infant cries, hyperphonation, weak vibrations and inspiratory phonations are present. However, these modes are not specific to any disease and presence of any of these modes does not guarantee the presence of any pathological state of the infant. This limits the use of spectrographic analysis for clinical applications.

In our experiments, it is observed that spectrograms get affected severely by additive noises such as white, babble, HR channel, vehicle noise, *etc.* (as shown in Figure 5.24- Figure 5.26). It is observed that the degradation of spectrogram resolution is less in the case of additive white noise, however, it is highest is HF channel noises at various SNR levels.

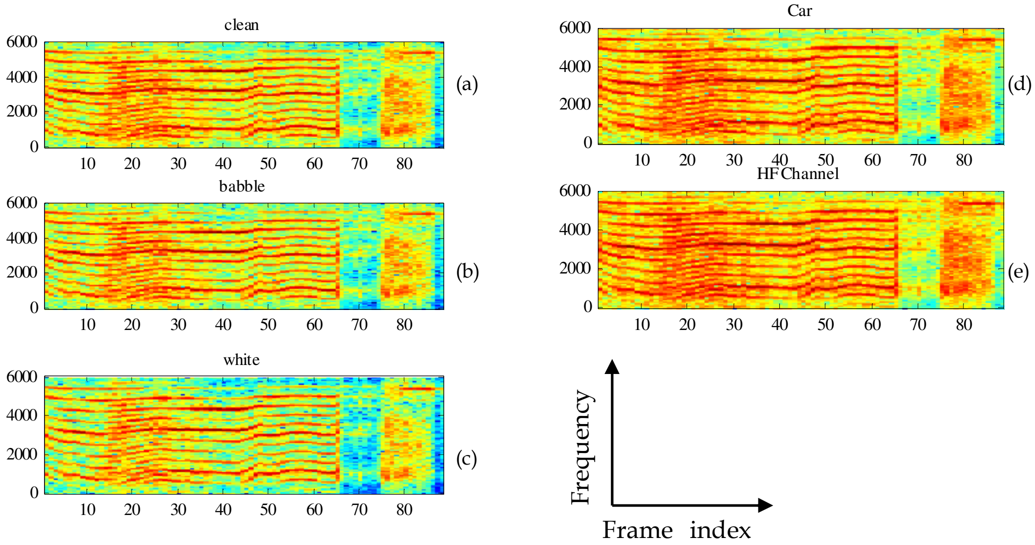


Figure 5.24. Effect of various additive noises on narrowband spectrograms of an infant cry at SNR level of (*10 dB*). Spectrogram for (a) clean signal, (b) babble noise, (c) white noise, (d) car noise and (e) HF channel noise.

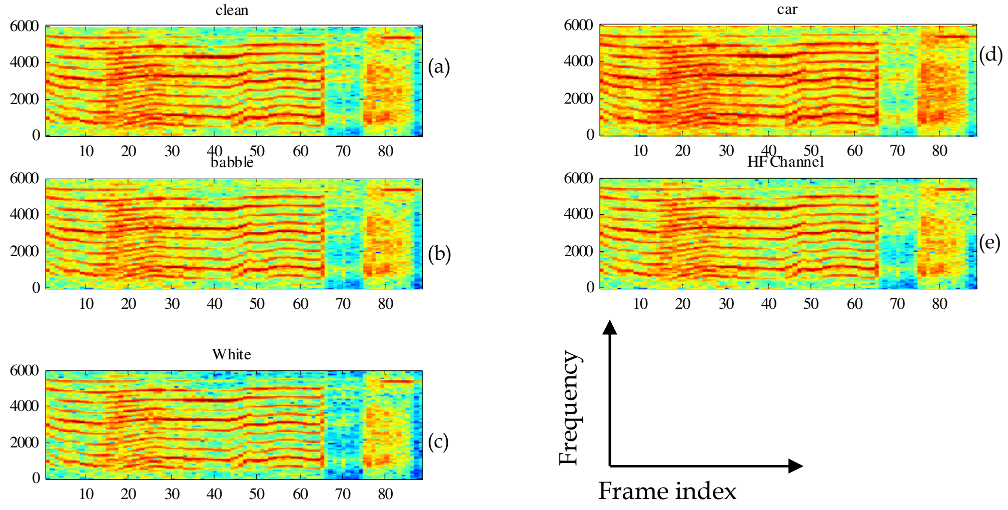


Figure 5.25. Effect of various noises on narrowband spectrograms of an infant cry at SNR level of (*5 dB*). Spectrogram for (a) clean signal, (b) babble noise, (c) white noise, (d) car noise and (e) HF channel noise.

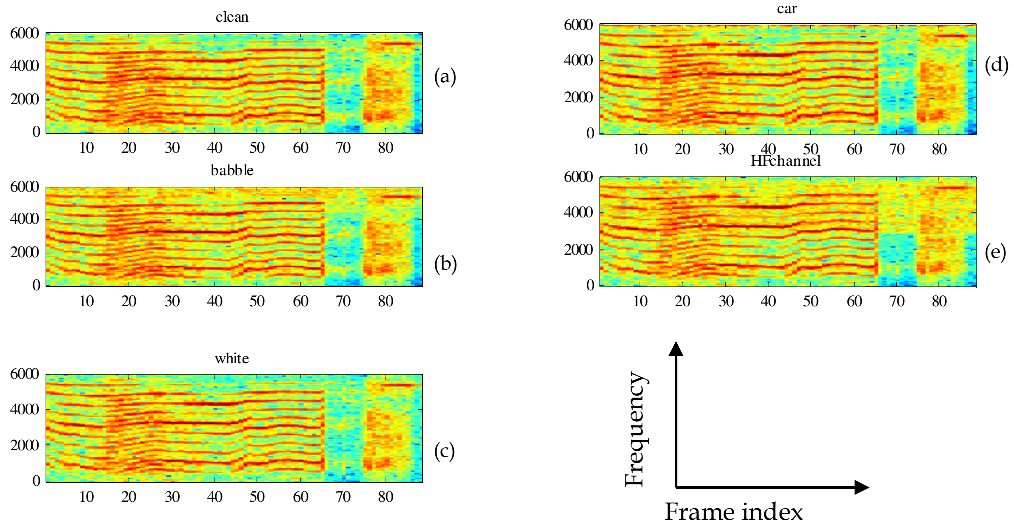


Figure 5.26. Effect of various noises on narrowband spectrograms of an infant cry at SNR level of (*-5 dB*). Spectrogram for (a) clean signal, (b) babble noise, (c) white noise, (d) car noise and (e) HF channel noise.

Though spectrographic analysis of infant cry can impart important information about the infants' health, yet its use is limited for research use only. Following are the limitations of spectrographic analysis:

a. Poor dynamic range and spectral resolution of the spectrogram.
b. Prior experience is required in spectrogram reading (and in addition, it is subjective and depends upon cognitive factors), and
c. Analyzing a large dataset with spectrograms is a tedious and time-consuming work.
d. Spectrograms employ a fixed duration window function $w(n)$ which limits the joint time-frequency atoms, *i.e.*, $w_{m,\omega}(n) = w(n-m)e^{j\omega n}$.
e. Spectrograms get effected by various additive noisy conditions as shown in Figure 5.24- Figure 5.26).

5.4 Classification of Adult Speech and Infant Cry Signals

For the spectral analysis of the adult speech and infant cries, the experiment is conducted on the Co*rpus II* and the features are extracted from the short-time

Fourier transform (STFT) of the signal. For the analysis of adult speech, TIMIT database is used. The sampling frequency of the recordings is *16 kHz*. In the TIMIT database, there are sentences spoken by the adult male and female speakers and the duration of these sentences is *1-2* seconds. The TIMIT database has training, testing and development sets given for the analysis of adult speech signal. From the training set, randomly selected *649* utterances and from the development set, *400* utterances of approximately *2-2.5 s* duration are selected. From this database, train and development sets are used for training and testing purpose. The infant cry *Corpus II* which has infant cries is also divided in train and test datasets. These infant cries are then divided in cryunits of duration (*0.75 sec- 1 sec* duration) to increase the statistical significance of the results. In the training set, there are *635* infant cryunits and in the test set, there are *155* infant cryunits.

Feature Extraction: Before extraction of the features, TIMIT database samples are downsampled to *12* kHz to make the sampling frequency same in both the corpus used in this analysis. From the STFT of the infant cry or adult speech signal, features are extracted. The features used in this study are the ratio of the distribution of energy in the *0-1* kHz, *1-3* kHz range, *3-5* kHz range and *5-6* kHz to the total energy of the spectrogram (*6* kHz being the maximum available bandwidth for F_s= *12* kHz, due to Shannon's sampling theorem). In infants dominant signatures of F_0 are present in less than *1 kHz* that motivated us to select the frequency band of *0-1 kHz* for the analysis)

$$E_{1n} = \frac{1}{M}[\sum_{\omega=0}^{1kHz} S(m,\omega) / \sum_{\omega=0}^{6kHz} S(m,\omega)], \tag{5.6}$$

$$E_{2n} = \frac{1}{M}[\sum_{\omega=1kHz}^{3kHz} S(m,\omega) / \sum_{\omega=0}^{6kHz} S(m,\omega)], \tag{5.7}$$

$$E_{3n} = \frac{1}{M}[\sum_{\omega=3kHz}^{5kHz} S(m,\omega) / \sum_{\omega=0}^{6kHz} S(m,\omega)], \tag{5.8}$$

$$E_{4n} = \frac{1}{M}[\sum_{\omega=5kHz}^{6kHz} S(m,\omega) / \sum_{\omega=0}^{6kHz} S(m,\omega)],. \tag{5.9}$$

where $m = 0,1,.. M-1$ (M is the total number of frames in the STFT) and ω is frequency and $S(m,\omega)$ is the spectrogram. The rationale behind using these subband energies is due to the fact that hearing is the process of detecting energy [128]. In the adult speech, features are extracted after the *0.5 sec* of duration to remove the effect of silence which is present in the beginning of the recordings and amplitudes are normalized between ± *1*. For both the classes, STFT is evaluated for the duration of the *0.5 sec*. The parameters used in the STFT computation are window length of *300* samples (*i.e., 40 ms*), with an overlap of *120* samples (*i.e., 10 ms*) and number of FFT points taken are *256*. These features are then analyzed for their significance in adult speech and infant cries. The results are reported in next Section.

5.4.1 Experimental Results

The experimental values of the features mentioned above are shown in Table 5.3 and Table 5.4. From Table 5.3 and Table 5.4, it can be observed that the energy in the infant cries lies above the *1* kHz frequency band, however, the distribution of energy is very low in the frequency band above *3* kHz. However, in the case of adult speech, it is minimum in the frequency band above *5* kHz. To observe the distribution of energy in infants and adults and to verify the fact that energy distribution in infants is low below *1* kHz, scatter plot is shown between features E_{1n} and E_{2n} for infants and adults (as shown in Figure 5.27). It can be observed that infant cry samples and adult speech samples are very well separated. However, separability of adult speech samples and infant cry samples are comparatively poor with features E_{3n} and E_{4n} as shown in Figure 5.27.

Table 5.3. Spectral energy features of adult speech

	E_{1n}	E_{2n}	E_{3n}	E_{4n}
Mean Values	0.38	0.34	0.13	0.15
Maximum Values	0.69	0.64	0.37	0.65
Min Values	0.12	0.05	0.03	0.01

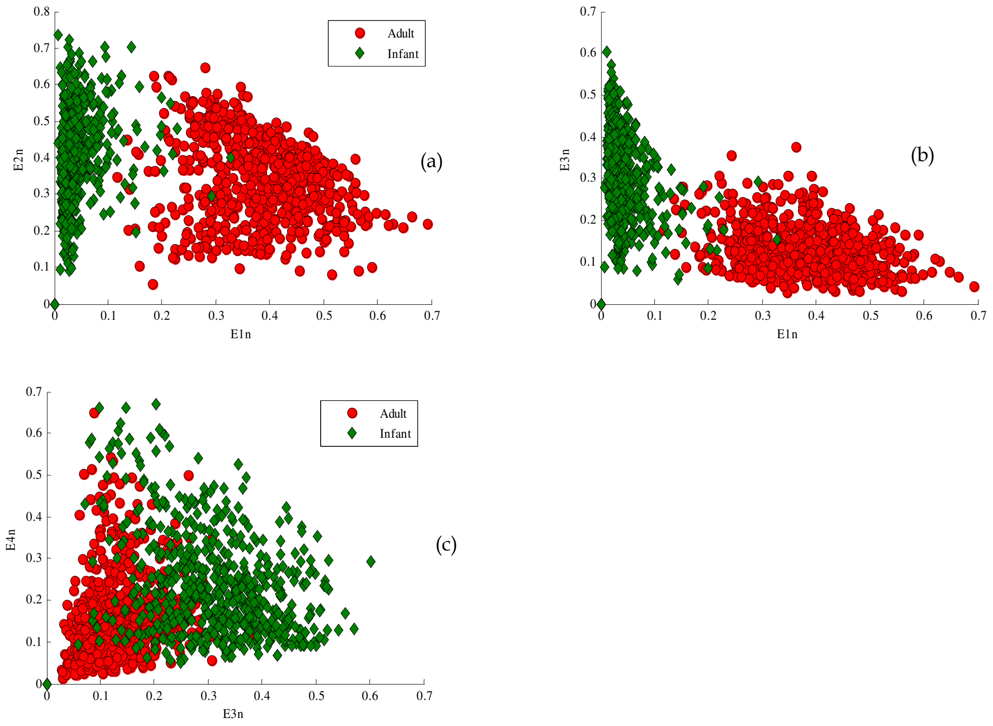


Figure 5.27. Scatter plots of features (a) E_{1n} and E_{2n} and (b) E_{1n} and E_{3n} (c) E_{3n} and E_{4n} derived for adult speech and infant cries. In all the subfigures, 'o' corresponds to adult speech and '+' corresponds to infant cry signal.

Table 5.4. Spectral energy features of infant cries

	E_{1n}	E_{2n}	E_{3n}	E_{4n}
Mean Values	0.04	0.40	0.31	0.23
Maximum Values	0.33	0.73	0.61	0.67
Min Values	0.01	0.09	0.06	0.05

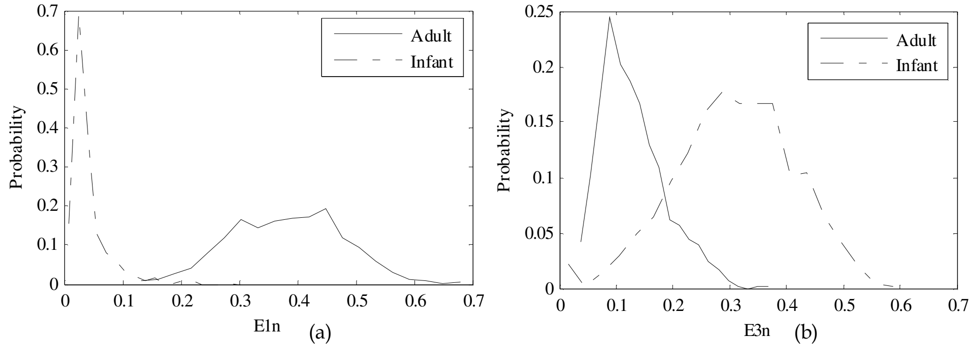


Figure 5.28. Probability density functions (pdf) of (a) E_{1n} and (b) E_{3n} for adult speech and infant cries.

Based on these results, these features are tested for their statistical significance using support vector machine (SVM) classifier. In this classifier, the radial basis function (RBF) kernel is used with kernel parameter γ=1. The classification results of *4*-fold cross-validation experiments are shown in Table 5.5. Classification accuracy (in %) is defined as the ratio of the correctly identified samples to the total number of samples in both the classes multiplied by *100*.

Table 5.5. Classification accuracy (in %) using spectral energy features for the classification of adult speech and infant cries

Feature	Fold-1	Fold-2	Fold-3	Fold-4	Mean
E_{1n}	98.56	98.13	97.68	98.52	98.22
E_{2n}	61.08	19.00	38.64	50.54	42.31
E_{3n}	91.53	87.18	75.55	82.55	84.20
E_{4n}	70.99	57.83	38.64	50.54	54.50

It can be observed from Table 5.5, that the energy distribution below *1* kHz is a significant feature in the classification of adult speech and infant cries. This is primarily due to the following reasons:

1. Differences in the length of the vocal tract (*i.e., 8 cm* in infants as opposed to *17.5 cm* in adults) results in shift in formants and their time-varying contours to lower side of the spectrum for adult speech than for infants or children.
2. Differences in F_0 (F_0 is in the range of *80* Hz to *250* Hz in adults as opposed to *250 Hz* to *500 Hz* (which could even go upto *1 kHz*) for infants.
3. Thus, in the case of infants, together with a shift in formants and dynamics in F_0 (during ten cry modes) and stressed vocal folds during crying, the resulting spectrum is having energy concentration in higher frequency region than to lower frequency counterpart (as clearly evident from Table 5.4).

Along with it, the distribution of energy is quite distinct in the *3-5 kHz* frequency band (as shown in Figure 5.38). In the *1-3* kHz frequency band, it is very difficult to find differences in the adult speech and infant cries. Same is observed for E_{4n} as well as can be observed from Table 5.5.

5.4.2 Conclusions

It has been observed that in infant cries, very little information is there in the below *1* kHz frequency band (primarily due to higher F_0 and hence, distantly spaced pitch harmonics). However, in adult speech, there is significant information in this frequency band. In the case of infant cries, the F_0 is high which is *3-4* times of the adult F_0. Hence, in the frequencies below *1* kHz, the pitch harmonics are comparatively low or sometimes absent (as in the case of hyperphonics cries). Thus, the energy in this band is very low. However, in high frequencies, the distribution of energy in both the class (adult *vs.* infant) is low. Hence, very high frequency bands (such as above *5 kHz*) cannot be used to classify these two classes. In the *3-5* kHz bands, of adult speeches very high harmonics are present and thus, they carry very less amount of energy. However, it carries *6th* or higher harmonics (for F_0 of *500* Hz) in infants. Thus, it carries high energy compared to adult speech (where this band contains *15*th or higher harmonics for F_0 of *200* Hz)

5.5 Analysis of Normal and Pathological Infant Cries

5.5.1 Preprocessing and Feature Extraction

In this analysis of infant cries, *corpus-I* is used. The data was collected at the sampling frequency of *44.1* kHz. From the cry samples, data is divided into cryunits. In all, we have *229* normal cryunits and *145* pathological cryunits. The cry signal is passed through a fourth order lowpass filter with cutoff frequency of *3 kHz*. Then, for each cryunit, voiced region is selected using energy measure (in particular, l^2 norm-based algorithm). From the voiced portion of the cryunit, F_0 contour is extracted using a sliding window of *30 ms* with overlap duration of *15 ms.* For each of the cry sound frame, mean fundamental frequency (F_0) is estimated using autocorrelation method. After finding the F_0 contour and the length of each cryunit (*i.e.*, duration), a feature vector (1×4) is formed. The feature vector is given by V= [min F_0, max F_0, mean F_0, duration]. For the estimation of F_0, autocorrelation method is used.

5.5.2 Autocorrelation Method

The method used for fundamental frequency estimation is the autocorrelation-based method [**148**]. In this method, the autocorrelation of the short segment of the signal is found. The autocorrelation of the signal $s(n)$ is given by [**148**]

$$R(l) = \sum_{n=0}^{N-1-l} s(n)s(n+l), \tag{5.10}$$

where l is lag element. The autocorrelation function is a non-invertible transformation of the signal which represents the structure of the waveform. Hence, for the pitch detection of a voiced segment of speech, if the signal $s(n)$ is periodic with period P, *i.e.*, $s(n) = s(n+P)$ then its autocorrelation function will also be periodic with the same pitch period P, *i.e.*, $R(l) = R(l+P)$. Using this property of the autocorrelation function, peaks of the autocorrelation function of cry signal are found. The difference of these peaks corresponds to the pitch of the signal. The fundamental frequency can be found from pitch using $F_0 = F_s(samples\ per\ \sec) / pitch\ period(samples)$. This method is illustrated in Figure 5.29.

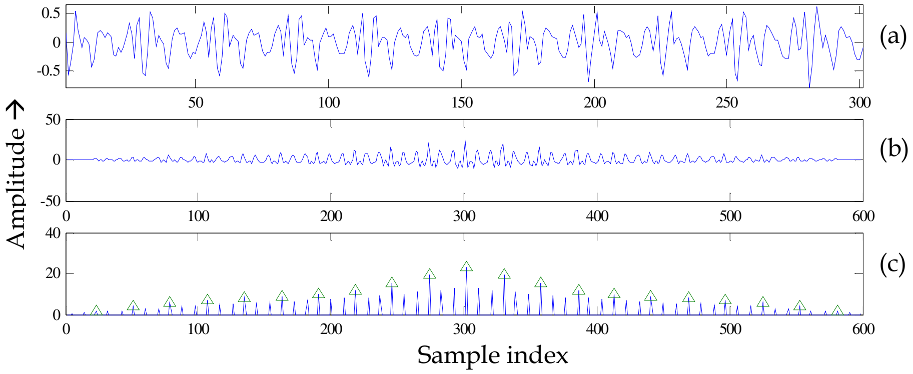


Figure 5.29. Estimation of pitch period using autocorrelation method. (a) time-domain infant cry signal, (b) its autocorrelation function and (c) peaks corresponding to pitch period. In all the subfigures, X-axis is sample index and Y-axis corresponds to the amplitude of the signal.

5.5.3 Experimental Results

One way ANOVA analysis [**149**] is conducted and significance of proposed features are observed for the following cases:

a. Normal *vs.* pathological cry
b. Normal pain *vs.* pathological pain cry
c. Normal pain *vs.* normal hunger cry
d. Pathological pain *vs.* pathological hunger cry

From Table 5.6, it can be observed that for normal pain cries, maximum F_0 is higher than the normal hunger cry. The mean F_0 of pathological cries is significantly higher than the normal infant cries and the same trend is observed in the pain cry analysis of normal and pathological cries. Minimum F_0 of pathological pain cry is higher than the hunger cry of similar type. Hunger cry is longer than pain cry in both normal as well as pathological infant cries. Minimum F_0 of pathological cries is higher than normal cries. Minimum F_0 of pathological cry is higher than normal pain cry. However, minimum F_0 of hunger cry in both cases are almost the same. Hunger cry of a pathological infant has higher mean F_0 than normal infant's hunger cry.

Table 5.6. Statistics of features for all cry types

Cry type ↓ / Features →	Minimum F_0 (Hz)	Maximum F_0 (Hz)	Mean F_0 (Hz)	Cry length (s.)
Normal pain (229)	206.89	765.14	415.07	1.48
Normal hunger (52)	216.53	737.02	416.38	2.14
Normal all (279)	206.82	779.24	405.44	1.60
Pathology all (145)	214.66	761.23	440.92	1.74
Pathological pain (52)	233.32	738.21	483.21	1.40
Pathological hunger (140)	216.87	763.54	437.27	1.87
*Numbers in brackets indicate number of cryunits				

Observations from Table 5.7, suggest that minimum, maximum and mean fundamental frequencies are good features for cry classification of normal *vs.* pathological cries. In both the cases, minimum F_0 and maximum F_0 are good features to identify hunger and pain cries as well. Duration does not seem to be a good feature for identifying the normal and pathological cries.

However, it is a good indicator of identifying pain and hunger cries in both normal and pathological infants. Mean F_0 feature does not serve as a feature for distinguishing normal pain and normal hunger cry (as the number of samples and number of classes in the analysis are same, it resulted in same values of degree of freedom 'Df' in the ANOVA analysis).

Table 5.7. ANOVA analysis of different cry types

		Df	Sd	F	P			Df	sd	F	P
Normal *vs.* pathological cry	**Min F_0**	452	14.34	32.07	2.64e-08	**Normal pain *vs.* normal hunger cry**	**Min F_0**	277	12.12	25.93	6.53e-07
	Max F_0	452	22.44	69.28	1.02e-15		**Max F_0**	277	37.40	23.18	2.42e-06
	Mean F_0	452	20.72	315.31	6.66e-54		**Mean F_0**	277	27.61	0.09	0.76
	Duration	452	1.246	1.45	0.22		**Duration**	277	1.23	11.97	0.0006
Normal pain *vs.* pathological Pain cry	**Min F_0**	279	13.28	172.71	4.94e-31	**Pathological pain *vs.* pathological hunger**	**Min F_0**	190	24.08	17.08	5.35e-05
	Max F_0	279	36.82	23.69	1.89e-06		**Max F_0**	190	32.62	23.25	2.89e-06
	Mean F_0	279	31.69	197.35	2.87e-34		**Mean F_0**	190	34.10	68.39	2.28e-14
	Duration	279	1.17	0.18	0.66		**Duration**	190	1.17	5.90	0.016
Df: Degree of freedom, sd: standard deviation, F: *F*-ratio, P =*p*-value (probability)											

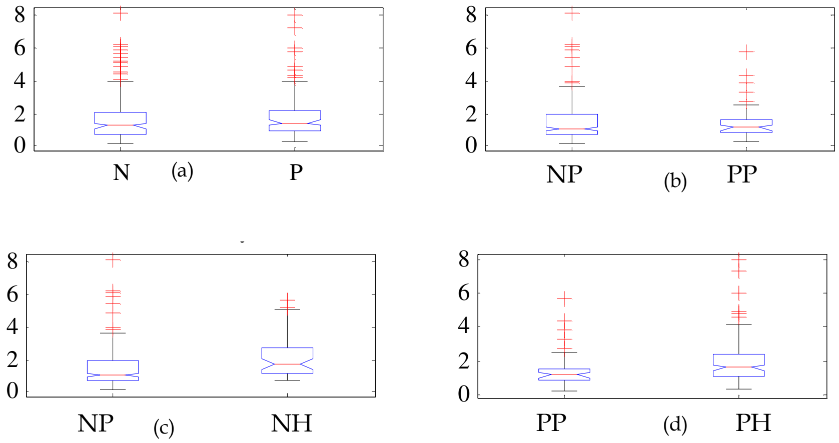


Figure 5.30. Box plots of duration features for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all subplots, Y-axis is time in sec. and X-axis represents the class type. **Notations**: N: Normal, P: Pathological, NP: Normal Pain, NH: Normal Hunger, PP: pathological pain, PH: Pathological hunger infant cry.

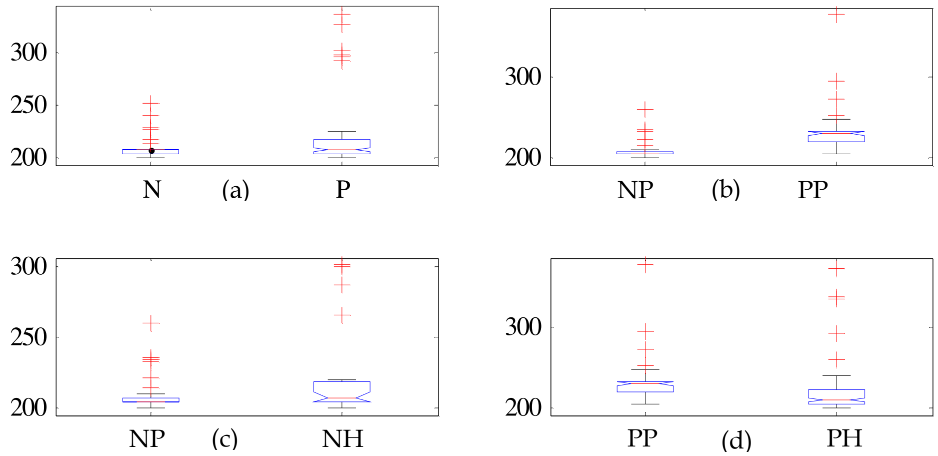


Figure 5.31. Box plots of minimum fundamental frequencies (F_0) feature for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is the frequency in Hz and X-axis represents the class type.

Looking at the boxplots shown in Figure 5.30 of durational features for all the four cases, it can be observed that the means of the two classes within each case are not much different. However, duration of hunger cries are found to be longer than the pain cries in both normal and pathological infant cries. Furthermore, normal pain cries are generally longer than the pathological cries. When minimum F_0 plots are analyzed as shown in Figure 5.31, it is observed that this feature is not very much different in the two classes in all four cases. Only pathological pain cries have higher minimum fundamental frequency than the normal pain cries

The mean fundamental frequency (F_0) of pathological cries is significantly higher than the normal infant cries and the same trend is observed in the pain cry analysis of normal and pathological cries as shown in Figure 5.33. There is no significant difference in the maximum F_0 of all four cases. Normally, pain cries have higher maximum F_0 (Figure 5.32). The significant difference in duration and F_0 of normal and pathological pain cries is the reason that these features can be used to classify normal *vs.* pathological cries.

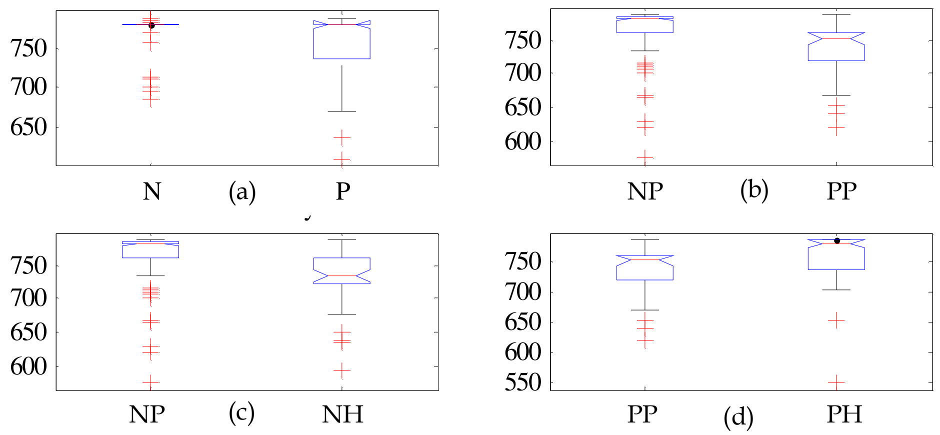


Figure 5.32. Box plots of maximum fundamental frequencies (F_0) feature for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is the frequency in Hz and X-axis represents the class type.

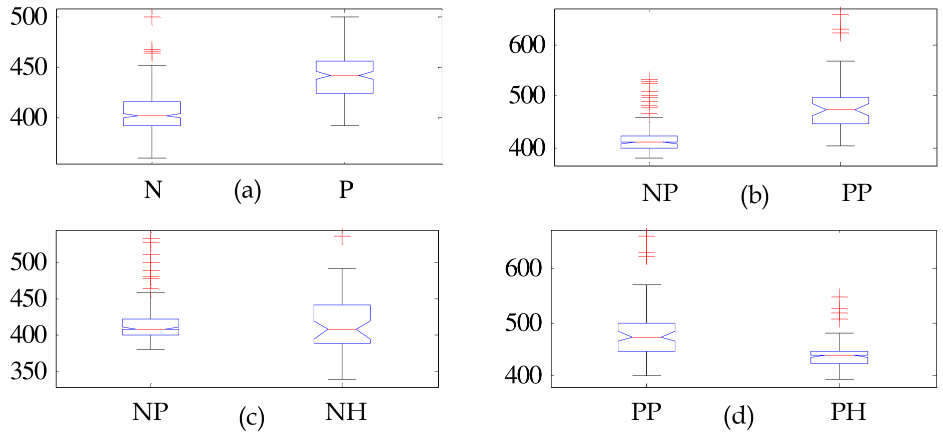


Figure 5.33. Box plots of mean fundamental frequencies (F_0) feature for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is the frequency in Hz and X-axis represents the class type.

Features derived from F_0 contour for a cryunit plays an important role in the characterization of cry type and identify the state of the infant. For hunger cries, generally the duration is longer and for pain cries, F_0 is higher. These features are important acoustic cues for a parent or a caretaker of an infant for recognizing the need of infant (hunger) and for identifying the urgency of cry call (in the case of pain cry). For infants who suffer from some pathology, it has been observed by their parents that their infant's cries are either shorter than normal infants or they have comparatively either higher or lower pitch than normal infants. Same is observed in our experiments as well. Change in F_0 of pathological cries can be attributed to instability in neural control of the larynx and lower vocal tract [**43**]. These parameters cannot be used alone for cry classification. However, these features can be used along with some suitable features for improving the results.

5.6 Newborn's Cry Analysis

As has been observed in the Section *3* that newborn cries are quiet distinct from the normal and comparatively mature infant cries. Some researchers have worked in the analysis of the first cry of the infants. Most of the work is done by the medical practitioners and researchers in this direction. In [**44**], authors have used larynx of two dead newborns to generate sounds by applying air pressure. Their finding shows that the role of the larynx is same as excised organ, free of neurologic control. Their role in the first cry is not to vibrate by themselves, however, to generate aerodynamic perturbations generating supraglottic vibrations. Complex interactions are responsible for the nonlinear phenomenon found in first cry signal. Neurological control and regulation is absent in the first cry. In another study, researchers have used the newborns' cries to find out the effect of prenatal exposure to cocaine [**31**]. In this Section, the distinction between first cry and other cry types is reported using different features and effectiveness of these features in infant development is presented.

Newborn is defined as the infant whose age is few days after the birth (< *1* month) or alternatively known as a *neonate*. However, the normal cry is the cry of infants older than the neonates. In this study, Corpus *II* is used. The distribution of the cries is shown in Table 5.8 and Table 5.9.

Table 5.8. Corpus statistics for infant cry analysis

Class	Number of infants
Newborn fullterm normal birth	45
Newborn pre-term	36

Table 5.9. Distribution of cry samples of newborn infant cries

S. No.	Type of cry	Age group	No. of participants	No. of samples
1.	Fullterm infant's birth cry	After birth	16	49
2.	Premature infant cry	< 20 days (mostly <10 days of age)	7	50
3.	Newborn hunger cry	<10 days age	16	30
4.	Newborn pain cry	<3 months (mostly <1 month)	17 (4 infants (6 cries) are older than 1 month)	19
5.	Newborn cry due to urination	< 20 days	12	24
6.	Newborn cry due to wet diaper	< 20 days	4	5
7.	Normal after birth	S. No. 3-6 combined		

BC: Birth cry, PC: Pre-mature infant cry, H: Hunger cry, P: Pain cry, U: Urination cry, W: Wet diaper cry, N: Normal Infant's cry

5.6.1 Estimation of Pitch (F_0) Contour

The autocorrelation method of the pitch estimation is widely used for pitch estimation in speech-related applications. In autocorrelation method of pitch estimation, the speech signal is divided into smaller duration frames because the speech is a non- stationary signal. For a frame of speech such as *20-30* ms (comprising of *2-3* pitch periods), after pre-processing which includes passing the signal through a lowpass filter, autocorrelation is found. Periodicity which is observed in the periodic signal is also observed in its autocorrelation function. The autocorrelation function is symmetric, the distance between two highest peaks is calculated which is equal to the pitch period of the signal. Autocorrelation method of F_0 estimation does not work well for infant cry analysis because in infant cry signals (due to surrounding noise), sometimes false (ambiguous) peaks are present which gives misleading false peaks and thereby, high frequency values (due to a decrease in the difference in two peaks). In this chapter, F_0 contour is estimated using modified autocorrelation method. In the pre-processing stage, the infant cry signal is passed through a 4th order Butterworth lowpass filter with a cutoff frequency of *1* kHz, in order to remove high frequency harmonics present in the signal. The filtered signal is then segmented into frames of duration *30* ms with an overlap of *50* %. On each of the cry signal frame, modified autocorrelation method is applied and peaks corresponding to the pitch values are identified and pitch is estimated. In modified autocorrelation method of pitch extraction, the signal $x(n)$ is clipped by a reference level C_L. The clipping level C_L is chosen as the *25* % of the maximum peak sample values. The resulting signal is given by:

$$y(n) = clc\left[x(n)\right] = \begin{cases} \left(x(n) - C_L\right) & , \quad x(n) \geq C_L \\ 0 & , \quad \left|x(n)\right| \langle C_L \\ \left(x(n) + C_L\right) & , \quad x(n) \leq C_L. \end{cases} \tag{5.11}$$

For the clipped signal $y(n)$, the autocorrelation function is found using the formula:

$$R'(m) = \sum_{n=0}^{N-1-m} y(n) \cdot y(n+m), \quad 0 \leq m \leq M_0, \tag{5.12}$$

where N is the length of the sequence, M_0 is the number of autocorrelation points to be computed, m is lag or delay. Clipping of the signal removes the added noise effects and hence, it performs better than autocorrelation method of pitch estimation. From the autocorrelation function applied to clipped signal, the peaks are identified. The difference of the peak locations, gives the estimate of the pitch of the signal. The examples of the modified autocorrelation method for pitch extraction applied to voiced and unvoiced segments of the cry signal are shown in Figure 5.34. We can observe that for unvoiced segments, the autocorrelation function have very less number of peaks and thus, the segments which have less than *6* number of peaks are taken as unvoiced segments and pitch is taken as zero for them.

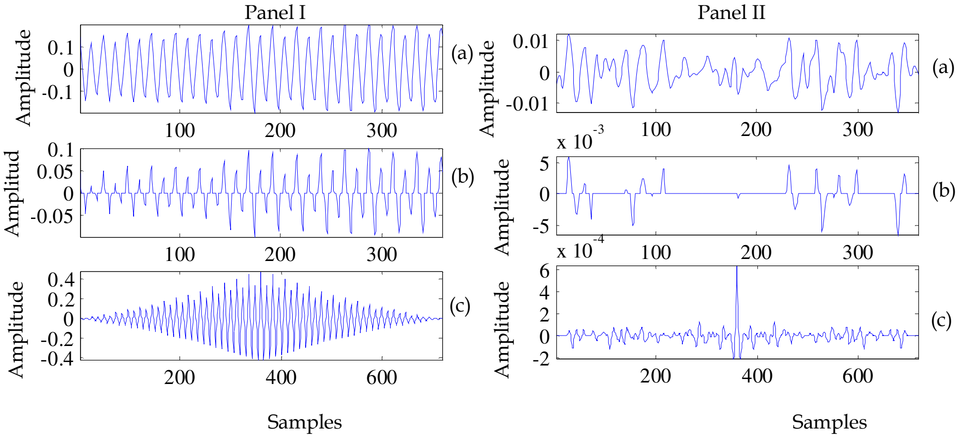


Figure 5.34. Modified autocorrelation algorithm for F_0 extraction for Panel I: voiced segments and Panel II: unvoiced frame. In all the subfigures (a) time-domain signal, (b) center clipped signal of (a) and (c) autocorrelation function of (b).

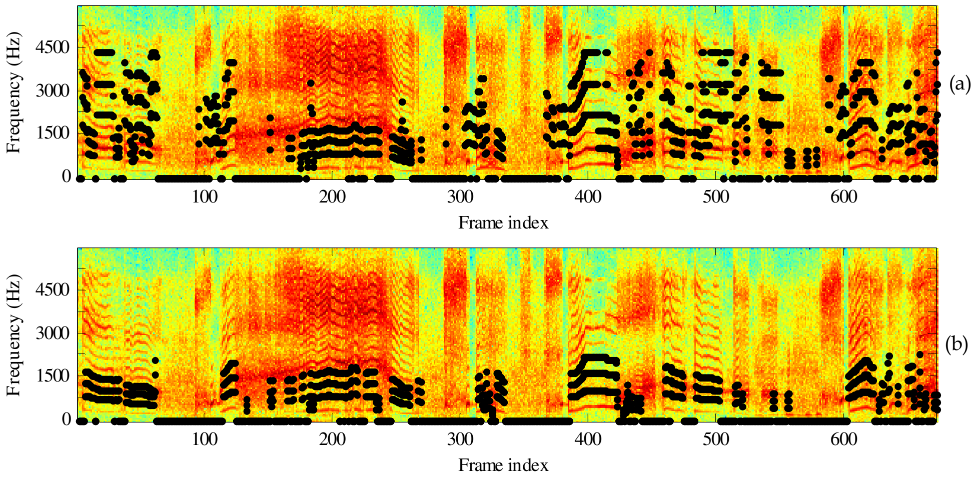


Figure 5.35. Comparison of pitch extraction methods (a) autocorrelation method (b) modified autocorrelation method.

In Figure 5.34 (Panel II), the modified autocorrelation method is illustrated for the voiced and unvoiced segments. In the proposed method, clipping-level was suggested as *64* % of the maximum peak amplitude. In the case of infant cry signals, it was observed by intensive computer simulation that keeping such a high threshold for clipping is removing most of the peaks of the signal, thereby does not work for pitch estimation. By the iterative method, we decided the threshold for clipping as *25* % and this is found to give best results for F_0 estimation. To compare the performance of the F_0 extraction with the standard autocorrelation method, spectrogram is used. In infants, reference glottal flow waveform (GFW) for comparing the performance of the F_0 extraction methods is not available. The glottal flow waveform cannot be acquired from the infants by non-invasive methods and hence, it limits the availability of the glottal flow waveform for infants. Thus, to compare the performances of the two F_0 estimation algorithms, we have used spectrogram. If the estimated harmonics match with the harmonics present in the spectrogram, we can say that the algorithm is better. This decision is made after observing the matching of harmonics with spectrogram for many infant cry samples in order to have a decision which is statistically significant. From Figure 5.35, it can be observed that the modified autocorrelation-based method of F_0 extraction works better than state-of-the-art method, *i.e.*, autocorrelation method of F_0 extraction.

5.6.2 Feature Extraction

It is known that our ears are sensitive to two parameters, namely, loudness and pitch. Loudness is associated with the amplitude of the signal, it is a perceptual feature which is recently found to be associated with the strength of excitation (SoE) [**150**]. However, the pitch is also a perceptual feature and is associated with the F_0 of the signal. Hence, to extract information of these two parameters, energy and F_0- related parameters are estimated and different cry signals are analyzed using them. For each of the cry sample, F_0 contour is calculated using the modified autocorrelation method and following features are estimated:

1. Minimum of F_0 contour
2. Maximum of F_0 contour
3. Mean of F_0 contour
4. Median of F_0 contour
5. Normalized energy of the signal (*E*)
6. Normalized energy in *0-2* kHz (*E1*)
7. Normalized energy in *2-4* kHz (*E2*)

8. Normalized energy in *4-6* kHz (*E3*)
9. Unvoicing percentage in the total cry (UV ratio)

The normalized energy of the signal is defined as the energy of signal divided by the length of the signal, *i.e.*,

$$E = \frac{1}{n} \| X(k,\omega) \|^2 , \tag{5.13}$$

where E is the normalized energy, n is the number of cry segments and $X(k,\omega)$ is the short-time Fourier transform (STFT) of the signal. The normalized energy of the signal is calculated for the three subbands, namely, (1) *E1*: *0-2* kHz, (2) *E2*: *2-4* kHz and (3) *E3: 4-6* kHz (because the data is recorded at *12* kHz sampling frequency and hence, the maximum available bandwidth is *6* kHz). The unvoicing regions are identified as the segments where the number of peaks in the autocorrelation function is less than *6*, thereby giving zero pitch frequency. The sum of frames with zero pitch values divided by the total number of frames present in the cry is considered as the unvoicing ratio of the cry signal.

Different cry types defined in Table 5.8 and Table 5.9 are analyzed using these features and analysis of variance (ANOVA) analysis is used to find the significance of these features in various infant cry types. The analysis and the results are given in next Section.

5.6.3 Experimental Results

Different cry features are analyzed for the reasons of crying of an infant for following cases:

a. Fullterm birth cry *vs*. premature newborn cry
b. Fullterm birth cry *vs*. newborn pain cry
c. Fullterm birth cry *vs*. newborn hunger cry
d. Newborn pain cry *vs.* newborn hunger cry
e. Newborn pain cry *vs.* newborn cry due to wet diaper
f. Newborn pain cry *vs.* newborn cry during passing the urine

g. Newborn cry due to wet diaper *vs.* newborn cry during passing the urine
h. Newborn hunger cry *vs.* newborn cry due to wet diaper
i. Newborn hunger cry *vs.* newborn cry during passing the urine
j. Newborn birth cry *vs.* newborn other reasons of crying (hunger or wet diaper or passing urine or pain).

Table 5.10. Mean values of the features for different infant cry types

S. No.	Type of cry signal	Min (F_0)	Max (F_0)	Mean (F_0)	Med. (F_0)	*E*	*E1 (0-2 kHz)*	*E2 (2-4 kHz)*	*E3 (4-6 kHz)*	UV ratio
1.	BC	167.86±18	561.16±38	436.22±57	417.70±46	0.487±0.2	0.167±0.05	0.245±0.12	0.078±0.04	0.844±0.07
2.	PC	169.80±20	564.95±16	429.31±46	416.42±35	0.344±0.17	0.169±0.07	0.119±0.07	0.057±0.04	0.687±0.14
3.	H	175.18±36	550.36±40	425.19±55	413.30±48	0.416±0.17	0.199±0.04	0.1543±0.09	0.066±0.05	0.693±0.14
4.	P	160.00±11	556.28±55	409.06±76	397.38±66	0.636±0.23	0.238±0.06	0.279±0.13	0.123±0.07	0.688±0.14
5.	U	173.91±34	552.62±48	387.48±74	388.45±63	0.435±0.15	0.198±0.05	0.168±0.08	0.070±0.04	0.674±0.21
6.	W	166.14±15	571.43±0	448.50±41	435.30±33	0.663±0.09	0.261±0.03	0.306±0.07	0.100±0.02	0.548±0.11
7.	N	170.51±30	553.85±45	411.15±67	403.19±57	0.491±0.20	0.212±0.05	0.198±0.11	0.083±0.05	0.677±0.17

BC: Birth cry, PC: Pre-mature infant's cry, H: Hunger cry, P: Pain cry, U: Urination cry, W: Wet diaper cry, N: Normal Infant's cry

The mean values of the above features along with the standard deviation are given in Table 5.10. For the simplification purpose, the analysis is taken separately for the F_0-based features and remaining features.

5.6.4 Analysis using Fundamental Frequency (F_0)-Based Features:

From Table 5.10 and Figure 5.36, it can be observed that the minimum F_0, maximum F_0 and median of F_0 are almost similar in all the cases. Thus, these features cannot be used to characterize or discriminate a particular infant cry type. However, mean F_0 feature is showing differences in some cry types such as newborn’s birth cry has mean F_0 of *436.22* Hz while this parameter is *411.15* Hz for the normal newborn’s cry. Differences in the hunger cry and pain cries of the newborns are also observed. In hunger cries, the mean value of the F_0 is *425.19±55* Hz, mean F_0 is *387.48±72* Hz for urination cries. Significant differences are not found in the different features of F_0, based on the reason of crying (except in the two cases mentioned above). In the birth cries as well, these features do not change with the gestation age (GA). These parameters are almost similar for a normal full term as well as for premature babies.

The ANOVA analysis of the parameters derived from the F_0 contour also suggests the similar results. The results of ANOVA analysis are given in Table 5.11 for all the features. Here, we have considered *95* % confidence interval in ANOVA analysis which means features which give *p*-value less than *0.05* are of significance in the analysis of those particular cry types.

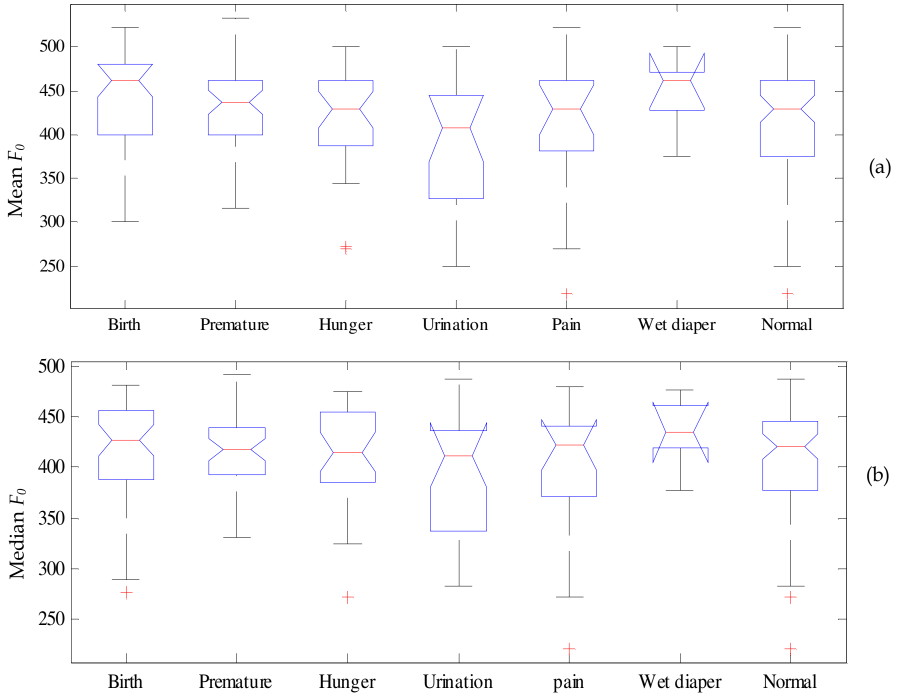


Figure 5.36. Boxplot for the F_0 features (a) mean F_0 and (b) median F_0.

Table 5.11. *p*-values obtained from ANOVA analysis of the newborn infant's cry.

S. No.	Class	Min F_0	Max F_0	Mean F_0	Med. F_0	*E*	*E1*	*E2*	*E3*	UV ratio
1.	Full term birth *vs.* premature	0.62	0.52	0.516	0.88	**0.0004**	0.85	**4.08e-8**	**0.0247**	**6.21 e-10**
2.	Full term birth *vs.* pain	0.092	0.68	0.117	0.16	**0.0129**	**6.84e-05**	0.34	**0.0016**	**2.29 e-7**
3.	Full term birth *vs.* hunger	0.24	0.24	0.41	0.69	0.1191	**0.017**	**0.0015**	0.256	**5.09 e-8**
4.	Pain *vs.* hunger	0.088	0.668	0.402	0.33	**0.0005**	**0.021**	**0.0004**	**0.0027**	0.91
5.	Pain *vs.* wet diaper	0.346	0.55	0.28	0.23	0.81	0.46	0.66	0.4962	0.06
6.	Pain *vs.* urination	0.107	0.82	0.35	0.65	**0.0015**	**0.03**	**0.0016**	**0.0042**	0.809
7.	Wet diaper *vs.* urination	0.633	0.404	0.089	0.13	**0.0036**	**0.0195**	**0.0023**	0.1372	0.22
8.	Hunger *vs.* wet diaper	0.59	0.26	0.392	0.34	**0.0047**	**0.0103**	**0.0029**	0.18	**0.04**
9	Hunger *vs.* urination	0.897	0.85	**0.039**	0.11	0.6792	0.97	0.58	0.70	0.70
10.	Newborn birth *vs.* normal	0.589	0.35	**0.035**	0.14	0.919	**4.29e-5**	**0.037**	0.57	**2.18e-9**

5.6.5 Analysis using Normalized Energy-based features

Analysis is done for various cry types mentioned in Section 5.6.3 based on normalized energy-based features. The mean values and standard deviations of the features are mentioned in Table 5.10.

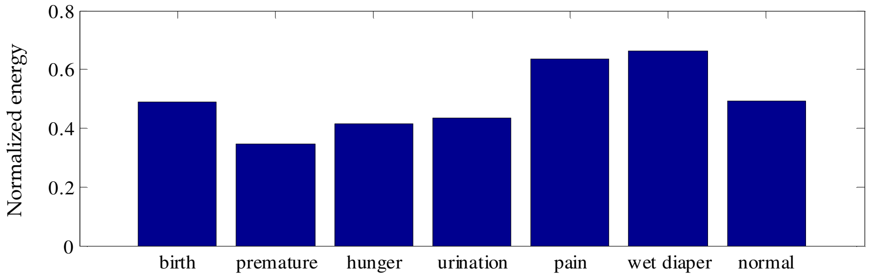


Figure 5.37. Barplot of mean values of normalized energy values for different cry types. Y-axis represents the normalized energy of the signal.

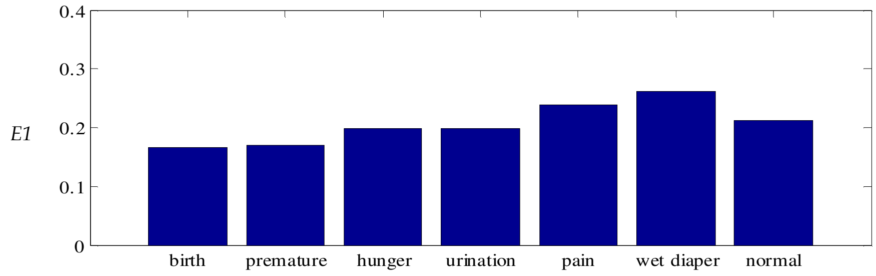


Figure 5.38. Bar plot of mean values of *E1 (0-2 kHz)* for different types of cries. Y-axis represents the normalized value of feature *E1*.

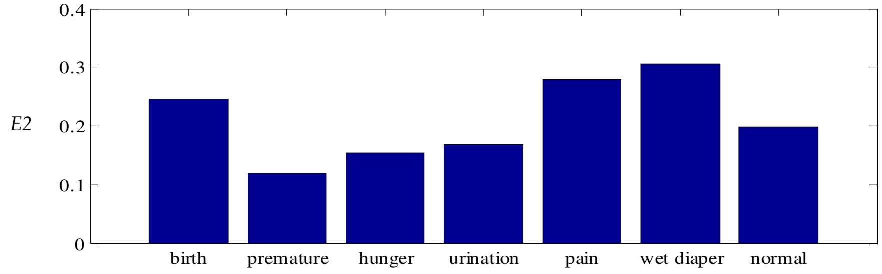


Figure 5.39. Bar plot of mean values of *E2* (2-4 kHz) for different types of cries. Y-axis represents the normalized value of feature *E2*.

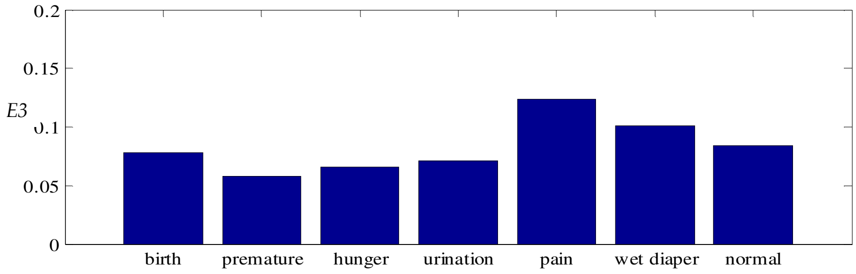


Figure 5.40. Bar plot of mean values of *E3 (4-6 kHz)* for different types of cries. Y-axis represents the normalized value of feature *E3*.

From Figure 5.38, it can be observed that normalized energy of the pain and wet diaper cries are higher than the other cry types. The energy is lowest in the premature infant cries. The energy of fullterm birth cries is higher than the premature infant cries. Comparing the distribution of the energy of the cry signals in the three frequency bands as shown in Figure 5.38 - Figure 5.40, we can observe that the pain cries and wet diaper cries have the highest energy in all the subbands. Moreover, most of the energy lies in the *2-4* kHz subband in all cries. In premature infants, distribution of energy is higher in lower frequency bands compared to normal full term infant birth cries (as shown in Figure 5.38), where the distribution of energy is higher in the mid frequency band (*2-4* kHz) (as shown in Figure 5.39). In hunger and urination cries, distribution of energy is more in lower frequency bands (*0-2* kHz) compared to pain and wet diaper cries where energy in *2-4* kHz band is higher. In the high frequency bands (*4-6* kHz), the distribution of energy is very low for infant's cries except for pain and wet diaper cries as shown in Figure 5.40.

Results of ANOVA analysis are shown in Table 5.11. It can be observed that the normal infant's birth cries are distinct from the premature infant's cries. Because of higher energy of normal full term infants, we can distinguish their cries from premature infants, who have low energy in the cries. The reason of cry can also be identified from the energy feature. Hunger cries are found distinct from the pain cries and wet diaper cries are found different from the crying while passing the urine. In the case of birth cry and premature infants' cries, it is observed that the energy difference is very high and this result in the identification of the cries by auditory analysis as well. The differences in the two cry patterns are there in the mid- frequency bands. In the band *2-4* kHz, the energy of the birth cry is higher than the pre-mature infant's cry and in other bands, the distribution of energy is same for both the cries. Birth cries of normal full term infants and pain cries are characterized by the high energy of the signal as shown in Figure 5.41 (a). ANOVA analysis in the three frequency bands shows that the two cry can be characterized by the distribution of energy in the low and high frequency bands. The energy is high in low and high frequency bands in pain cries compared to birth cries as shown in Figure 5.41 (b) and Figure 5.41 (d).

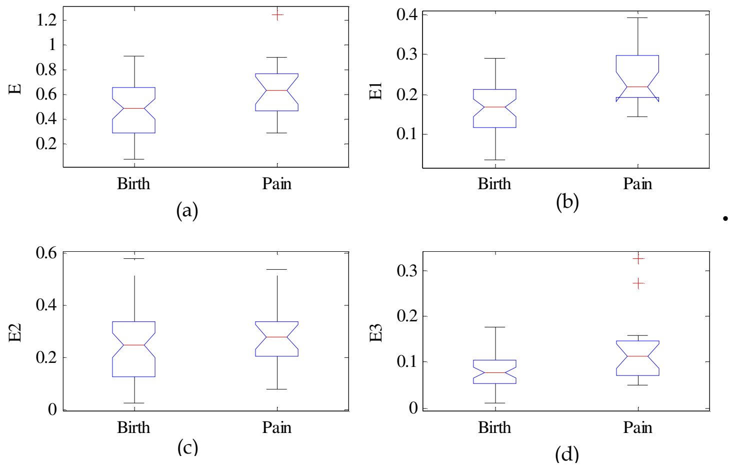


Figure 5.41. Boxplots of normalized (a) *E*, (b) *E1*, (c) *E2* and (d) *E3* for birth and pain cries.

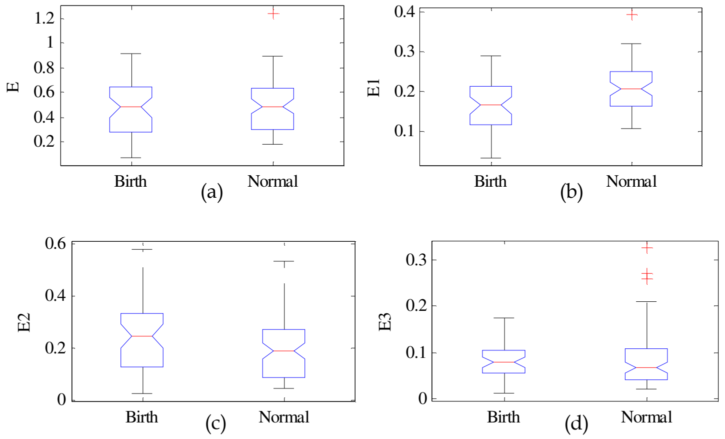


Figure 5.42. Boxplots of normalized (a) *E*, (b) *E1*, (c) *E2* and (d) *E3* for birth and normal cries.

Analysis of hunger, pain, wet diaper and urination cries shows that distribution of energy is similar in hunger and urination cries as well as in pain and wet diaper cries. These two groups of the cries are distinct from each other on the basis of total normalized energy as well as energy in their respective bands. However, it is difficult to characterize differences in hunger and urination cries using energy-based features. Similar is the case for the classification of pain and wet diaper cries, where the energy in all the bands is almost similar irrespective of the reason of crying. Normal fullterm birth cries are different from the other reasons of crying such as hunger, pain, wet diaper and urination named here as normal cry, on the basis of *E1* and *E2*. In birth cries, *E2* is higher than the other reasons of crying. However, in normal crying (due to other reasons of crying) energy *E1* is higher than birth cries of fullterm healthy infants as shown in Figure 5.42.

5.6.6 Analysis Using Unvoicing Ratio of the Cry

From Figure 5.43 and Table 5.10, we can observe that the birth cries are characterized by very high unvoicing ratio. Compared to cries due to hunger, pain, wet diaper and urination, this higher unvoicing ratio makes birth cries distinct from other cry types. This feature is found to be useful in classifying the reason of crying also where energy-based features are not working. Similar energy-level of cries can be classified according to the ratio of crying present in the cry. Pain and wet diaper cries which have similar energy in all the frequency bands can be distinguished by using UV ratio analysis. In pain cries, UV ratio is higher than the wet diaper cries. Similarly, between wet diaper and hunger cries, hunger cries are found to have more unvoicing and can be distinguished from cries due to a wet diaper.

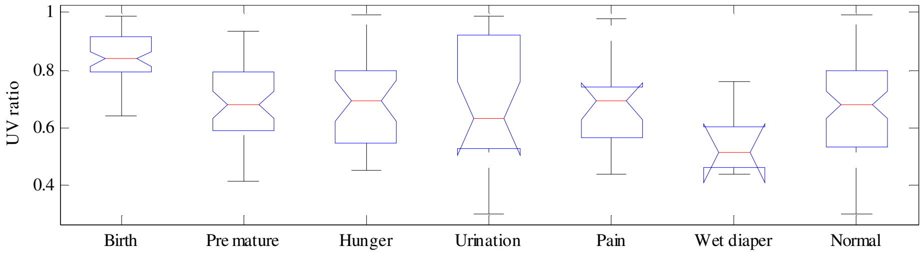


Figure 5.43. Boxplot for the UV ratio in the infant cries.

5.6.7 Analysis of Results

In this Section, newborn infants cries are analyzed for the various reasons of crying such as hunger, pain, wet diaper and while passing the urine. These are the various reasons of crying in a newborn. For the analysis of the cries, features used are the F_0-based features, energy-based features and the unvoicing ratio of the cry segments. Some important results from the above analysis are as follows:

a. Birth cry can be characterized by high energy and high unvoicing ratio. The reason for this is, as soon as the newborn come in contrast with the external world from the mother's warm womb; it is his or her response to the external stimulation. At the time of birth, there is poor regulation of central nervous system (CNS) over vocal folds working. At birth cry, lungs open up for the first time and breaths air instead of sack fluid [**11**].
b. Most of the energy in birth cry is located in the frequency band *2-4* kHz. However, normal infant's cry is having its maximum distribution of energy in *0-2* kHz (*i.e.*, normal, hunger, urinating). Pain cry is also having the same characteristics of having higher *E2* than *E1*.
c. Compared to other infant cry types, pain cries and wet diaper cries have higher energy distribution in *4-6* kHz frequency range. Higher energy in higher frequency ranges asks for the attention of the care taker and informs that a quick action is required. In the other words, higher frequency content in the cry reflects the urgency of the attention and discomfort to the infants.
d. Characteristics of hunger cry and cry during passing the urine found to be similar to all the parameters. Similarly, pain cries and wet diaper cries have similar characteristics.
e. Hunger cry and cry during passing the urine can be distinguished from each other using mean F_0 parameter. Remaining parameters are the same for them.
f. Unvoicing ratio in infants is an indicator of the maturity of infant's vocal production system. In birth cry, high unvoicing indicate that, in the birth cry, vocal folds movement is very irregular which results in poor voice quality of the cry. With the production of the birth cry, infant's neural system integrates and within few days, cries become rhythmic.

g. Wet diaper cries can be distinguished from the pain cries based on the feature of unvoicing ratio. In pain cries, it is found to be higher than wet diaper cries.

h. Mean F_0 in newborn birth cries is higher than the normal infant's cries. There are no significant differences in the birth cries of newborns and those of premature infants cries. This indicates that until infant achievs a minimium gestation age (GA), vocal folds do not vibrate to produce voiced cry sounds.

i. F_0 - related features are not useful in identifying the reason of crying in newborns, though it is a useful parameter in infant (who is more than *1* month of age) cry analysis for understanding the reason of cry.

5.7 Significance of Unvoiced Segments and Fundamental Frequency (F_0) in Infant Cry Analysis

In this Section, the significance of unvoiced segments and F_0 in infant cry analysis is investigated. To find out the unvoiced segments from the infant cry, F_0 contour is used. For extraction of F_0 contour, Teager Energy operator (TEO)-based pitch extraction algorithm is used. As discussed in Section 4.6, TEO gives the running estimate of the signal energy in terms of its amplitude and instantaneous frequency (as opposed to traditional short-time l^2 energy measure used in signal processing literature). To quantify the importance of proposed features in infant cry ANOVA method is applied. It has been found that quantification of unvoiced segments and fundamental frequency in the cry, deliver information about the maturation of cry production system. In infant cry analysis, the presence of high unvoicing ratio in a cry cannot be attributed to the presence of pathology as it happens in adult vocal fold pathological sounds.

5.7.1 Teager Energy Operator (TEO)-Based F_0 Estimation Algorithm

In this Section, F_0 estimation from speech signals [**151**] is explained. In the pre-processing stage, the signal is passed through a *4*th order lowpass Butterworth filter with cutoff frequency of *1* kHz (because highest F_0 in infants is upto *1* kHz). In addition, it was observed by the Teagers [**128**] that the bumps within two consecutive GCIs are suppressed tremendously, if we bandpass filter speech to lower frequency region and thus, it avoids interference of large peak (due to sudden closure of glottis) with these bumps (due to nonlinear aspect of speech production). The filtered signal is divided into smaller duration frames. For each of this speech frame, TEO profile of the signal is calculated. Computation of TEO requires three consecutive samples of speech as shown in eq. (4.25).

The TEO profile of a signal gives the running estimate of its energy. From the TEO profile of the signal, the peaks of TEO profile are identified. The difference in peak location gives a rough estimate of the pitch period. The pitch for a speech frame is calculated as follows:

a. Calculate the number of peaks in a frame. If the number of peaks is more than three, then take median of the value as pitch period of the frame, else take the frame as a unvoiced segment. Median of the pitch period is considered to ignore the effect of outlier pitch period (due to the unvoicing effect) instead of mean as proposed in [**151**].
b. After getting the pitch period for each of the frame, calculate F_0 by dividing the pitch period from sampling frequency.

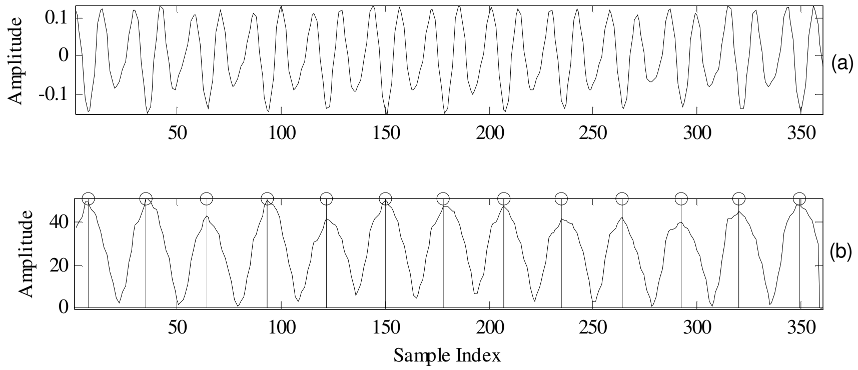


Figure 5.44. Extraction of pitch values from TEO profile. (a) cry signal and (b) TEO profile of (a). Adapted from [**152**]

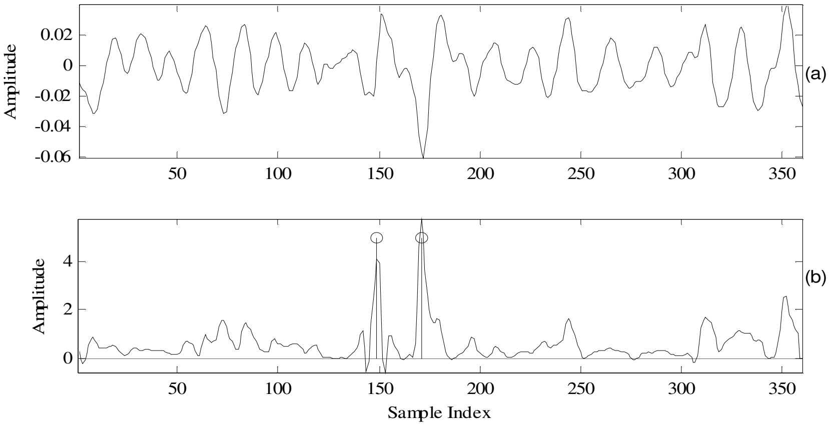


Figure 5.45. Extraction of pitch values from TEO profile (unvoiced signal). (a) cry signal and (b) TEO profile of (a).

From Figure 5.44, it can be seen that TEO signal is able to pick the peaks in the voiced signals. It is interesting to note that the bumps are not present in the TEO profile of the bandpass filtered infant cry signal (as opposed to strong bumps observed in fullband infant cry signal shown in Section 4.6, Figure 4.20). The unvoiced segments are considered as the segments for which the F_0 is zero (algorithm cannot find three consecutive peaks) as shown in Figure 5.45. Other algorithms of pitch estimation do not perform well in peak-picking because of high variability in pitch values and selection of a threshold for peak-picking. This algorithm has already been tested for adult speech signal [**151**].

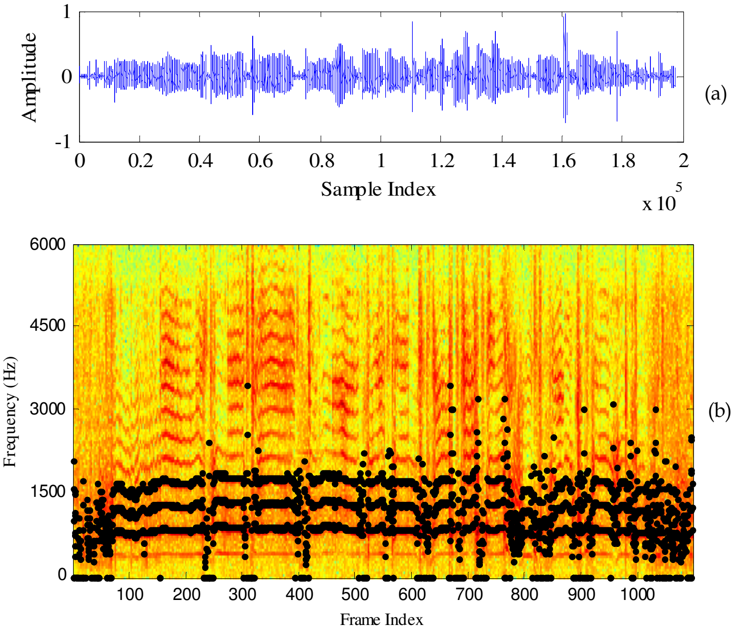


Figure 5.46. Spectrogram of infant cry: (a) infant cry signal and (b) spectrogram and fundamental frequencies (F_0) contour extracted from TEO (a), Y-axis represents NFFT bins which corresponds to frequency in Hz). Adapted from **[152]**.

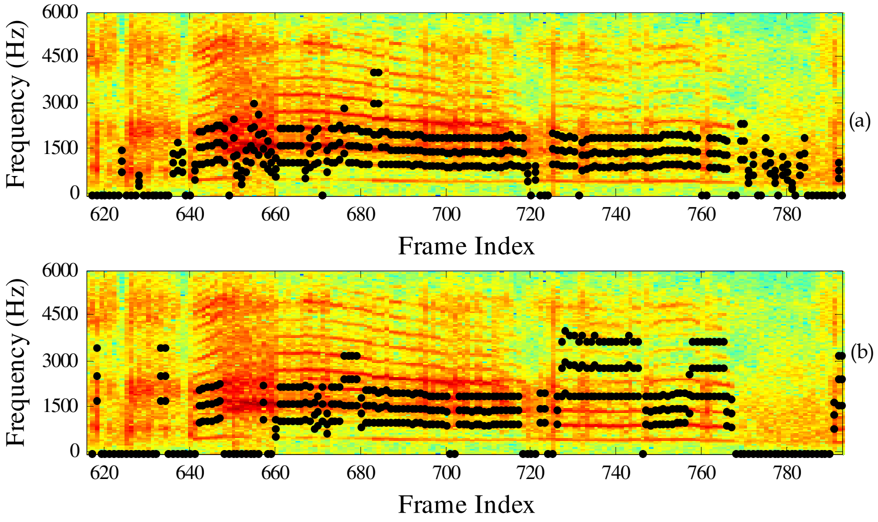


Figure 5.47. Comparison of fundamental frequencies (F_0) estimation algorithms (a) TEO-based and (b) autocorrelation-based method.

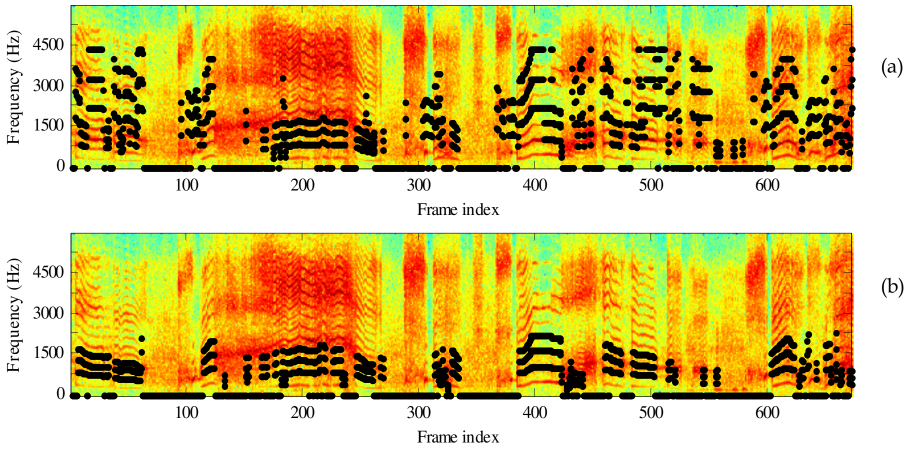


Figure 5.48. Comparison of fundamental frequencies (F_0) estimation algorithms (a) TEO-based and (b) modified autocorrelation-based method.

Box 5.1. MATLAB code for plotting fundamental frequencies (F_0) contour on the spectrogram.

```
% code for plotting F0 contour on spectrogram

[teo_f0] = f0_teo(s,fs);% find F0 using TEO method
% find F0 using modified autocorrelation method
[f0_corr] = f0_autocorr_modified(s2,fs);
% plot spectrogram of the signal
Xspec= spectrogram(s2,300,120,256,1E3);
% plot of spectrogram
AX(1)=subplot(2,1,1);
imagesc(flipud(log(abs(Xspec))));
hold on
% plotting f0 values on the spectrogram
plot(129-round(2*(teo_f0*129*2/fs)),'K.');
hold on
plot(129-round(3*(teo_f0*129*2/fs)),'K.');
hold on
plot(129-round(4*(teo_f0*129*2/fs)),'K.');
axis tight
AX(2)=subplot(2,1,2);
imagesc(flipud(log(abs(Xspec))));
hold on
plot(129-round(2*(f0_corr*129*2/fs)),'K.');
hold on
plot(129-round(3*(f0_corr*129*2/fs)),'K.');
hold on
plot(129-round(4*(f0_corr*129*2/fs)),'K.');

linkaxes(AX,'x');
axis tight
```

Because of unavailability of the ground truth for verification pitch extraction algorithm, F_0 contour and its harmonics are plotted on the spectrogram as shown in Figure 5.47. Comparison of TEO-based F_0 extraction method with autocorrelation method and modified autocorrelation method is also given in Figure 5.47 and Figure 5.48. Autocorrelation method of pitch extraction has been explained in Section 5.5.2. The correctness of this method is illustrated in Figure 5.47 whereas MATLAB code for this task is shown in Box 5.1. As can be observed from the Figure 5.47, at many points, F_0 is misidentified (between *720-760* samples). It can be observed that the TEO-based method is able to track small changes in F_0 (between samples *725-740*).

5.7.2 Experimental Setup and Results

The percentage of unvoiced segments is found over the complete cry utterance. In this piece of work, the significance of such unvoiced regions in infant cry signal is elaborated for infant cry analysis.

Database: In our experiments, *Corpus II* is used. Details of number of cries for normal, newborns and pathological cries considered in the experiment are given in Table 5.12 (as discussed earlier in Chapter 3).

Table 5.12. Number of infant cries for different classes

Infant Health Condition	No. of infant cries	Average duration of cries
Normal	90	43 sec
Newborn	40	42.6 sec
Asthma	7	41.7 sec
HIE	16	28.25 sec
Meningitis	4	29 sec
Fits (epilepsy)	4	51.25 sec
Misc. (heart disease, broken bones, larynx not developed, jaundice, cleft lip, high risk)	10	43.8 sec

In the pre-processing stage, the infant cry signal is passed through a 4^{th} order Butterworth lowpass filter with cutoff frequency of *1* kHz. After the signal is being filtered, the filtered signal is divided into frames of *30* ms duration with *15* ms of overlapping. For each of the frame, the pitch period is found using TEO-based F_0 extraction algorithm. After finding the percentage of unvoiced frames, which corresponds to the frames where minimum three peaks are not available. The values of the unvoicing ratio are shown in Figure 5.49.

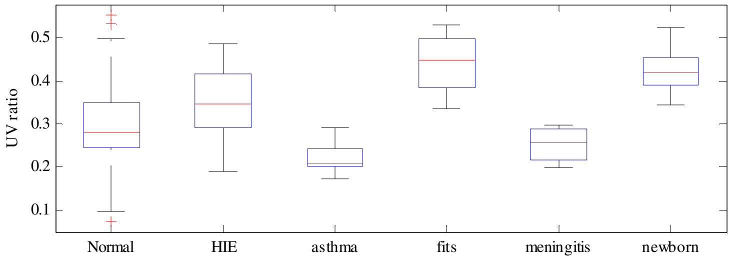


Figure 5.49. Boxplot of the unvoicing ratio in infant cries.

Table 5.13. ANOVA analysis of infant cry signal for significance of unvoiced frames

S. NO.	Case	F-Ratio	Probability
1.	Normal *vs.* HIE	5.64	0.019
2.	Normal *vs.* fits	11.67	0.001
3.	Normal *vs.* pathology HIE/meningitis/fits	7.39	0.007
4	Normal *vs.* Newborn	84.04	1.22e-15
5.	Normal *vs.* Asthma	4.78	0.032
6.	All	22.81	4.893e-17

The unvoiced segments are tested for their significance in infant cry analysis using ANOVA analysis. Results are shown in Table 5.13. It can be seen that the value of *F*-ratio is much higher than *1* and hence, it suggests the rejection of the null hypothesis that all infants came from the same group. This implies that relative amount of unvoiced regions is a significant feature in the analysis of infant cry. From Table 5.13, we can see that the *F*-ratio of unvoiced frames in a cry is higher in pathological infants cries compared to the normal infant cries. It shows the change in the unvoicing-to-voicing ratio in an infant due to the presence of pathology. Same results are obtained for normal infants and newborn (birth cry) analysis. In neonates as well, the percentage of the unvoiced segments are different than the normal infants. Details of unvoiced frames to total cry frames ratio is given in Table 5.14.

Table 5.14. Mean values of unvoiced frames in a cry episode. Adapted from [**152**]

S. NO.	Infant class	Mean UV	Standard Deviation
1.	HIE	0.3467	0.0848
2.	Meningitis	0.2518	0.0443
3.	Fits	0.4396	0.0810
4	Normal	0.2920	0.0846
5.	Newborn	0.4224	0.0444
6.	Asthma	0.2212	0.0390

The mean values and standard deviation of the unvoiced ratio given in Table 5.14, indicates that fits and HIE cries have a higher number of unvoiced frames. Asthma and meningitis have comparatively higher voiced frames. This observation suggests that asthma cry is similar to a normal cry. Normal infant cries have a wide range of unvoicing ratio because it consists of all cries of infants of all age groups from neonatal to *1* year old. Moreover, in pain cries, other studies in the literature have reported higher values of unvoiced regions in spectrograms of the cry. In neonates, the percentage of unvoiced frames is much higher than their normal counterpart and it is even higher than pathological infant cries. It has already been observed that in adult pathological speech, the unvoiced segments are higher compared to healthy adult speech. Hence, a feature which can capture this unvoicing information may perform better in pathological and healthy speech classification.

Table 5.15. ANOVA analysis of infant cry signal for significance of F_0

S. NO.	Case	F-Ratio	Probability
1.	Normal *vs.* asthma	15.82	0.0001
2.	Normal *vs.* fits	2.93	0.09
3.	Normal *vs.* pathology HIE/meningitis/fits	0.08	0.782
4	Normal *vs.* Newborn	21.96	7.22e-06

Table 5.15 shows the F-ratio values for ANOVA analysis carried out for finding the significance of F_0 in infant cry analysis. It can be observed that for *95 %* confidence interval, the mean F_0 feature does not seem to work well for normal and pathological infant cry analysis. The *F*-ratio values are significantly higher for normal and newborn birth cries and normal and

asthma infant cries, *i.e.*, asthma infants and newborns have different values of F_0 or different functioning or anatomy of vocal folds.

5.7.3 Analysis of Results Obtained

For an infant, crying is a way to express his or her emotions and the needs. Production of cry requires coordinated functioning of respiratory, laryngeal and supralaryngeal network and neurophysiological coordination. Early research work on infant cry analysis is based on spectrographic analysis of infant cry. Researchers have reported many cry modes from these spectrographic modes. The primarily used modes are phonation, hyperphonation, dysphonation, double harmonic breaks and glide. Identifying these modes from the spectrogram requires expertise and experience. To identify the phonation and dysphonation in the spectrogram, proposed work in this thesis can be used. It can be seen from Figure 5.50 that the proposed method coincides with the harmonics present in the spectrogram and also able to identify dysphonation regions. These regions are addressed in this work as unvoiced frames. The presence of dysphonation or unvoiced frames indicates the presence of noise and nonlinearity in the cry production system. In adult voice pathology, the presence of this noise represents dysfunction of vocal folds. Similarly, a higher percentage of unvoicing in cry represents less coordination among cry production organs or neural integration in infants.

Table 5.16. Mean values of F_0 in a cry episode

S. NO.	Infant class	Mean F_0 (Hz)	Std. (Hz)
1.	HIE	331.04	53.15
2.	Meningitis	397.42	19.54
3.	Fits	333.33	0
4	Normal	354.07	65.58
5.	Newborn	280.8	24.8
6.	Asthma	443.75	46.1

On the other side, higher unvoicing in newborn birth cry indicates less integration of cry production mechanism. In newborns, vocal folds are not fully developed and hence, the newborn birth cry lack phonation in its spectrogram as shown in Figure 5.50. Integration of vocal system and development of vocal folds in the first three months of infant life has been reported in [**153**]. This work quantifies the noise reduction in cry production with infant age during initial *3* months of age.

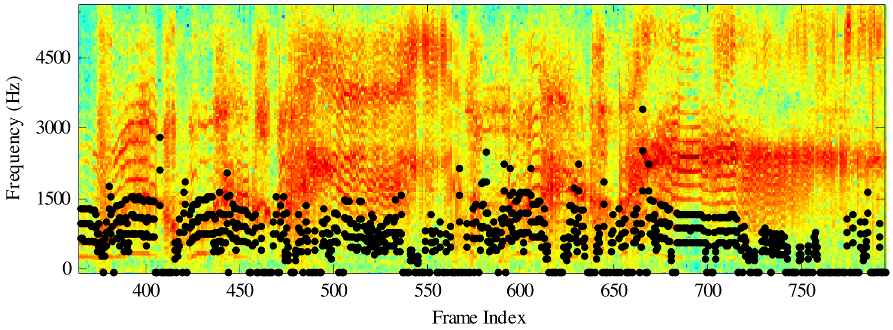


Figure 5.50. Spectrogram of newborn birth cry. X-axis is frame index and Y-axis represents FFT bins which corresponds to frequency (*128* bin= *6000* Hz).

Apart from the percentage of unvoicing in the cry, analysis of F_0 using ANOVA is shown in Table 5.15. It can be seen that mean F_0 does not carry a significant role in normal and pathological cry classification except in classifying normal and fits (epilepsy) infant cries. The mean values and standard deviation of the F_0 vales are shown in Table 5.16. We can observe that the mean F_0 of normal and newborn cries are significantly different. This observation also suggests the same result mentioned earlier, that with growing age infant vocal folds are developed and cry production mechanism becomes comparatively mature [**153**]. In the case of fits, which is a neurological disorder, can be a cause of change in F_0 compared to the normal infants and presence of high unvoiced regions in entire cry episode.

5.8 Sudden Infant Death Syndrome (SIDS)

Sudden infant death syndrome (SIDS) is the condition of infant's death where the reason of death remains unanswered even after thorough medical examination and autopsy. However, sudden unexpected death in infancy (SUDI) or sudden unexpected infant deaths (SUID) refer to deaths in infancy where the reason may be explained or unexplained. The distinction between SIDS and SUID is generally very difficult. Most of the SIDS deaths occur during the *1-3* months of age. The chance of deaths due to SIDS reduces after *1* yrs of age. It is also observed that most of the SIDS deaths occur in cold weather. Among other factors responsible for SIDS are mothers who are of less than *20* years of age, prenatal exposure to cigarette, tobacco and nicotine. The prone or side sleep position increase the risk of rebreathing expired gases, resulting in hypercapnia and hypoxia. This position also increases the risk of overheating by a decrease in heat loss and increasing body temperature compared to normal infants. The risk for SIDS is exceptionally high for infants who sleep on their stomach. This interesting observation led to a breakthrough investigation by studying the sleep postures of infants residing in East Germany *vs*. West Germany when Berlin wall was broken. Side sleeping is recommended only in exceptional cases for infants with upper airway disorder for whom the airway protective mechanism are impaired, which may include infants with anatomic abnormalities, such as type *3* or type *4* laryngeal clefts, who have not undergone antireflux surgery [**154**]. Premature infants are more likely to be at risk for SIDS compared to normal infant groups. Pre-mature infants should be placed in spine position for sleeping as soon as possible after the birth. Other recommendations to avoid SIDS are as follows:

1. The crib should be of safety approved.
2. Infant should not be allowed to sleep on thr sofa or soft beds.
3. Bed sharing with parents is not recommended in specific cases such as where parents are using toxic drugs, alcohol and cigarettes.
4. Car seats and other sitting devices are not recommended at home for routine sleep.
5. Wedges and positioning devices are not recommended.

6. Avoid alcohol and illicit drugs during pregnancy and after infant's birth.
7. Breastfeeding is recommended.
8. Swaddling does not reduce the chances of sleep. However, if it is applied to an infant who can roll, it can increase the risk of SIDS. There are insufficient evidences to show that swaddling should be used in routine to calm the infants. If it is correctly applied, this may avoid hazards such as hip dysplasia (misalignment of the hip joint) and strangulation.
9. Infant should be immunized as per the recommendation of hospital authorities.
10. Infants should sleep on their back instead of sleeping on their stomach.

In India, the infant mortality rate (infant deaths per *1000* live births) is *38* which is much higher than other developed countries (*3* in Australia and *6* in USA in 2011-15) [**155**]. Following the above recommendations can reduce the risk of SIDS to an infant.

5.8.1 Cry Characteristics of SIDS Victims

It has been reported by the parents of the SIDS victims that the cries of their infants are strange or different than the siblings and other normal infants. Stark and Nathanson studied the cries of a male infant who died at the age of *6* months. They found that the cries are shorter and weaker compared to the normal infants [**94**]. Colton and Steinschneider reported the cry characteristics of a female infant who died at the age of *63* days and reported that the F_0 was lower, cry duration was longer and sound pressure level (SPL) was higher than the normal infant's group and SIDS siblings group [**83**].

The results reported by the studies are contradictory. Thus, it is difficult to say that identification of the SIDS prone infants on the basis of some parameters is possible. The study on SIDS is very difficult because enough statistically meaningful data for cry analysis is not available to have statistical confidence during analysis of results and it is not known that the cries of SIDS infants are normal or abnormal with respect to characteristics embedded in their cries.

In a study reported by National Institute of Health (NIH), it is found that the structural difference in a specific part of the brain which causes risk for SIDS [**94**]. In a study reported by Harrison on SIDS infants, where he removed the larynges of the *74* infants who died of SIDS, it was found that the SIDS can be attributed to decrease in the subglottic area (around the age of *3* months), which is highly dangerous. The reduction in the subglottic airway is often secondary to an increasing mucus secreting glands, caused by upper respiratory tract infection [**156**]. Thereby, result in changes in the acoustic features of the cry.

Since enough data of SIDS infants is not available and making such database is not feasible, some infant groups are identified to study about SIDS infant's cry characteristics. These are (1) SIDS siblings and (2) high risk infants (neurological disorders that are found in SIDS victims) who are prone to SIDS.

5.8.2 Respiratory Instability and SIDS Siblings

Prolonged sleep apnea is of significance in the study of SIDS. Infants who have recurrent and prolonged sleep apnea were studied. The apneic pauses were longer and frequent in such infants. A measure of respiratory instability, namely, *PSA-4* score, is also found low in this group. *PSA-4* score is defined as [**84**],

$$PSA-4=-2.695+0.607(MT)+0.023(AR)+0.042(AN)-0.143(A/D), \qquad (5.14)$$

where an apneic pause is defined as the cessation of breathing for at least *2* seconds and *MT* is the mean duration of all apneic pauses, *AR* is the percentage of *REM* (Rapid Eye Movement) epochs during which at least one apneic pause was initiated, *AN* is the percentage of *NREM* (*i.e.*, non-rapid eye movement) epochs during which at least one apneic pause was initiated and *A/D* is 100 × the summed duration of all apneic pauses divided by the duration of the sleep.

It has been found that infants with high respiratory instability have lower mental and psychomotor development. Colton and Steinschneider in *1980* conducted a study to find the relationship between the cries of the infants and the respiratory instability in infants. Infants considered in this group were fullterm infants, premature infants and SIDS siblings. All infants were studied in the first and fourth week of their life [**157**].

The cries were analyzed based on the following acoustic features namely, fundamental frequency, sound pressure level (SPL), duration of cry, center frequency of first formant (F_1), center frequency of second formant (F_2), center frequency of the third formant (F_3), energy in the frequency band *50* Hz to *4* kHz, energy in frequency band *4-8* kHz and energy in *8-16* kHz. The mean value of F_0 is highest in the normal infants group compared to premature and SIDS siblings. There is no difference in the durational feature of the three groups. Center frequency of the first formant is found highest (*1600 Hz)* in the SIDS siblings, which is even higher than the theoretical values of the first formant. The higher value of the first formant is because of the wider opening of the mouth while crying. The sound pressure level in the *4-8* kHz band is also found highest in the sibling group of infants. Centre frequencies of all the three formants are higher in the case of sibling groups. In the fourth week cries, the average F_0 of the normal infants was found to be lower than the siblings and premature infants. The variance of the features in the fourth week of the age is smaller than the first week of the life.

There is greater respiratory instability in infants at risk for SIDS. The reason for respiratory instability is unknown. However, this occurs when the infant is sleeping. To understand the relationship among cry, respiratory instability and development, one needs to understand the relation of all these systems with CNS. CNS dysfunction can be one of the reasons of developmental delay in the SIDS siblings and respiratory instability. Respiratory instability may be because of the lesion on some part of the respiratory system. Interestingly, very recently a study supported by NIH, USA, found that infants who suffered from SIDS have an abnormality in the medulla oblongata of brain stems which is known to control the breathing function. In another study, it was found that the SIDS siblings have a significantly high pitch (F_0) compared to healthy term infants [**85**].

5.9 Analysis of Infant Cries of Respiratory Distress and Infants at Risk

In this Section, infants at risk who are suffereing from respiratory distress (and hence, possibly prone to SIDS) are analyzed. The feature used for the analysis is the signal's short-time energy. The short-time signal energy reflects the amplitude variations and is defined as

$$E(i) = \frac{1}{N} \sum_{n=1}^{N} | x_i(n) |^2 , \qquad (5.15)$$

where N is the length of the signal. For the computation of short-time energy, Hamming window of *512* points duration is chosen with *50* % overlap. The cry samples of infants suffering from respiratory distress are collected from *10* newborns (*i.e.*, *1* recording per infant). In the category of high risk *1* cry recording of brain hemorrhage, *2* cry recordings of Down's syndrome, *2* cry recordings of hydrocephalus, *2* of malnutrition (worst case scenario), *3* cry recordings of pyromeningitis, *5* of siblings of infants who have history of deaths in infancy and *3* cry recordings of infants suffering from thalesimia major was considered. All these cries are divided into smaller cryunits and for each of the cry l^2 energy is calculated and then ANOVA analysis is applied. The data considered in this analysis is from *Corpus I* and *Corpus II*. Cry samples of *Corpus I* are downsampled to *12* kHz before passing it through a *4*th order Butterworth lowpass filter of *5* kHz. The statistics of the energy values in the three classes are given in Table 5.17 and the distribution of normalized energy values is given in Figure 5.51.

Table 5.17. Values of energy features for high risk, respiratory distress and normal infants' cries

Class	Min E	Mean E	Max E	Number of cryunits
High Risk	9.21	108.51	283.25	52
Respiratory Distress	1.52	10.13	54.85	55
Normal	0.3894	35.56	211.88	792

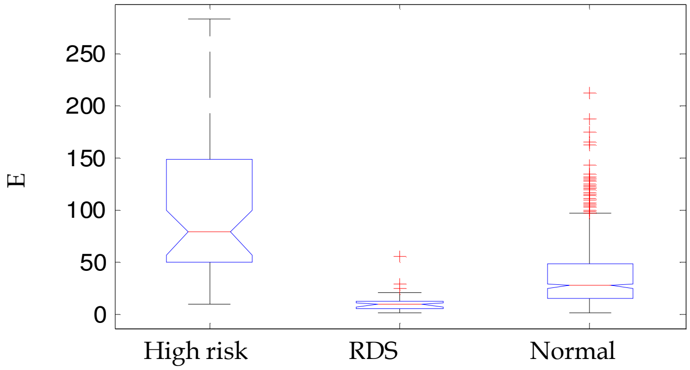


Figure 5.51. Bar plots of energy in the infants with high risk, RDS and normal health condition.

ANOVA analysis gives *F*-ratio value as *125.48* and *p=9.06e-49*. The high value of *F*-ratio ensures that energy of the signal in the three classes is a significant discriminative feature. In the high risk infants, the mean energy of the cry signal is found to be higher than the normal infants while in infants suffering from respiratory distress the energy of the infant cry signal is lower than the normal infants because of the problem in breathing.

Though the relationship among the different cry characteristics is not known and very less is known about the cry production mechanism in infants, the cry analysis in the first week of life may predict the development in later ages. Irregular changes in the cry acoustic may also indicate the sickness of the infant and timely diagnosis of the diseases may help in saving healthy lives and reducing the chances of metal and physical developmental delays.

5.10 Chapter Summary

In this work, spectrographic analysis for the infant cry analysis is explained and applied to analyze various infant cry types. It is shown that spectrographic analysis performs poor in the analysis of infant cries in order to identify the pathologies. Thus, because of limitations of spectrograms, automated algorithms are required for the analysis and classification of infant cries. In this chapter, F_0 extraction algorithm using modified autocorrelation and TEO is proposed for infant cry pitch (F_0) extraction. These algorithms work well for infant cries and can also be used for defining spectrographic modes from pitch (F_0) contours, automatically. It was shown that higher unvoicing percentage (*i.e.*, dysphonation) cannot guarantee the pathological state of an infant, as it is also visible in newborn birth cries and neonatal cries. The F_0 values also changes with the maturation of cry production system as observed from normal and newborn cry analysis. Significant changes in pitch from normal infant cries are indicators of some pathological state of an infant. It is observed in some genetic diseases, *e.g.*, cri-du-chat, where F_0 is different from normal infant cries. The significance of energy features is also confirmed using ANOVA analysis of the normalized energy feature. It has been found that in birth cries, most of the energy lies in *2-4 kHz* range. In the case of urgent calling signals (cries) issued by the infant, such as pain and wet diaper, the cry signal carries higher energy in the high frequency components which draws the attention of the care taker. In case of hunger cries, if the infant is prevented from the feeding from long duration then his or her hunger cries carries similar characteristics of the pain cry, because it conveys the more discomfort status of the infant compared to a normal hunger cry (which is generated at regular intervals) as reported in [**158**]. The characteristics of cries of normal and pathological cries are found different based on acoustic features.

Analysis of infants who are more at the risk of SIDS is also shown and it is observed that theses infants have cries which have either higher or lower energy than the normal infant's cries. In infants suffering from respiratory distress, energy in the cry is very low because of the problem in breathing. In the next chapter, an attempt is made to classify normal and pathological infant cries.

Chapter 6.

Classification of Normal and Pathological Infant Cries

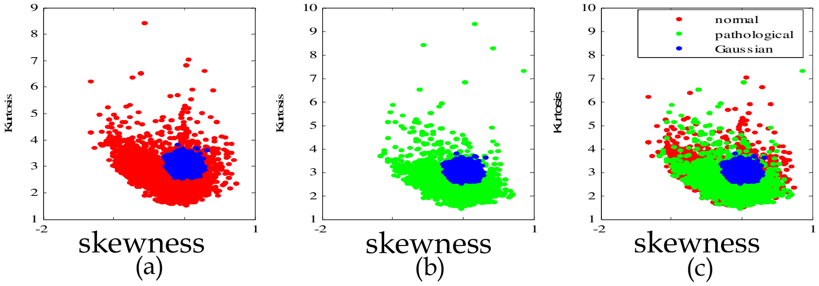
6.1 Introduction

In adult pathological speech a lot of work has been done towards classification, analysis of voice disorder and development of sytem for disordered speech [**159**], [**160**], [**161**], [**162**]. Signal processing methods have also been developed to aid medical science [**163**], [**164**]. However, there is need to develop some methods for infant cry analysis and classification. In this chapter, higher-order spectral analysis (HOSA) is applied on infant cry signals for classification of normal infant cries from pathological infant cries. From the family of higher-order spectral analysis, bispectrum is considered for the proposed task. Bispectrum is the spectrum of the third-order cumulant function. To extract features from the bispectrum, application of higher-order singular value decomposition (HOSVD) theorem is proposed.

Classification of pathological infant cries is presented as well. The work done in this direction is towards the classification of cries of infants suffering from asthma with those suffering from Hypoxy Ischemic Encephalopathy (HIE). To the best of author's knowledge, this has never been reported using any computer-based algorithm. For the classification of asthma and HIE infant cries, four different feature extraction approaches based on modulation spectrogram, glottal inverse filtering, auditory spectrogram and group delay are used and their performances are compared in this thesis. Another set of pathological cry classification considered in this thesis is the classification of normal and deaf infant cries.

6.2 Higher-Order Spectral Analysis (HOSA)

It is very common practice to use power spectrum-based features for various signal processing tasks. The power spectrum of a signal indicates the distribution of energy among frequency components at the penalty of losing phase information. The power spectral analysis gives information which is contained in the autocorrelation function of the signal under consideration. This description of signal works well when the signal under study is a Gaussian signal because for such signals, the higher-order cumulants are zeros (since Gaussian functions are described completely by only first two moments, *i.e.*, mean and variance). In realistic scenarios, the signal distribution may not be Gaussian and it necessitates the use of the higher-order description of the signals such as higher-order spectral analysis. To illustrate that the infant cry signal is not Gaussian, Figure 6.1 shows the distribution of *skewness* and *kurtosis* features of the four classes, namely, normal cries, pathological cries, Gaussian signals (of the same length and almost same in number to the normal infant cry samples) and adult voices. For a Gaussian signal, skewness is around zero. It can be observed from Figure 6.1 (c) that the infant cry samples do not have zero skewness and hence, we can infer that it is not a Gaussian signal. Similar results are found in the adult speech data signal as well. In Figure 6.2, skewness *vs.* kurtosis parameters are shown for the adult speech signal. In an adult speech as well, it is observed that the skewness is not zero, which shows the non-Gaussian nature of the speech signal and correctness of the HOSA application for speech analysis. The pathological voice signals are taken from the standard Massachusetts Eye and Ear Infirmary (MEEI) database [**165**]. Another important observation from Figure 6.2 is that with the vocal fold-related pathologies, deviation of skewness from zero increases. This suggests the need of HOSA techniques (bispectrum) for normal *vs.* pathological voice. On the other hand, it is observed from Figure 6.1 (a)-(c), that for infant cry data, it is very difficult to classify normal *vs.* pathological cries based on only skewness and kurtosis parameters (because the regions of normal and pathological infant cries are overlapping in Figure 6.1).



(a) (b) (c)

Figure 6.1. Distribution of skewness and kurtosis features for (a) normal and Gaussian signals, (b) pathological and Gaussian signals and (c) normal, pathological infant cries and Gaussian signals.

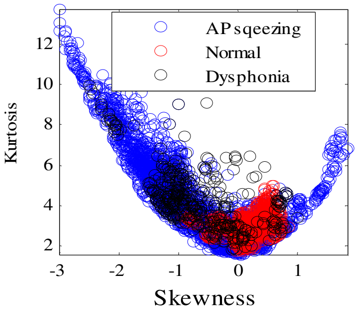


Figure 6.2. Skewness and kurtosis feature distribution for adult voices.

Higher-order spectra are defined in terms of their higher-order statistics, *i.e.*, cumulants of a signal. The third-order spectrum is known as *bispectrum* and the fourth-order spectrum is known as *trispectrum* of a signal. The power spectrum is also from the family of higher-order spectra. It is second-order spectra which is derived from the Fourier transform of the autocorrelation function (also known as Weiner-Khinchin theorem). Higher-order spectra and statistics are defined in terms of moments and cumulants. The autocorrelation function is the second-order cumulant function. For a Gaussian signal, the higher-order cumulants (of order more than 2) are zero which results in almost zero values of higher-order spectrum (as shown in Figure 6.2). Moments and moment spectra are useful in the analysis of deterministic signals while for stochastic signals cumulants and cumulant spectra are of importance. Speech is a stochastic signal and hence, higher-order spectra can give a better representation of speech signals. The motivations behind the use of HOSA are following:

1. To detect deviations from Gaussianity,
2. To identify and reconstruct non-minimum phase signals,
3. To suppress additive Gaussian noise and
4. Detect and characterize nonlinear properties in signals and identify nonlinear systems [**166**].

The n^{th} order moment function of a signal $X(k)$ is defined as

$$m_n^x(\tau_1,\tau_2,...,\tau_{n-1}) = E\{X(k)X(k+\tau_1)...X(k+\tau_{n-1})\}, \tag{6.1}$$

where $\tau_1,\tau_2,...,\tau_{n-1}$ are the time differences and $E\{.\}$ denotes the statistical expectation. The n^{th} order cumulant function of a non-Gaussian signal is given by:

$$c_n^x(\tau_1,\tau_2,...,\tau_{n-1}) = m_n^x(\tau_1,\tau_2,...,\tau_{n-1}) - m_n^G(\tau_1,\tau_2,...,\tau_{n-1}), \tag{6.2}$$

where $m_n^x(\tau_1,\tau_2,...,\tau_{n-1})$ is the n^{th} order moment function of signal $X(k)$ and $m_n^G(\tau_1,\tau_2,...,\tau_{n-1})$ is the n^{th} order moment function of an equivalent Gaussian signal that has the same mean and autocorrelation sequence as that of $X(k)$. Using cumulant definition, power spectrum, bispectrum and trispectrum are defined as follows [**166**]:

$$Power\ \ Spectrum:\ P(\omega) = \sum_{\tau=-\infty}^{\infty} c_2^x(\tau)\exp(-j(\omega\tau)), \qquad (6.3$$

$$Bispectrum:\ \ B(\omega_1,\omega_2) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} c_3^x(\tau_1,\tau_2)\exp(-j(\omega_1\tau_1+\omega_2\tau_2)), \tag{6.4}$$

$$Trispectrum:$$

$$C(\omega_1,\omega_2,\omega_3) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty}\sum_{\tau_3=-\infty}^{\infty} c_4^x(\tau_1,\tau_2,\tau_3)\exp(-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)), \tag{6.5}$$

$|\omega_1|<\pi, |\omega_2|<\pi, |\omega_3|<\pi$. For bispectrum, $|\omega_1+\omega_2|<\pi$ and for trispectrum, $|\omega_1+\omega_2+\omega_3|<\pi$. Excellent description of properties of higher-order spectrum analysis is given in [**167**].

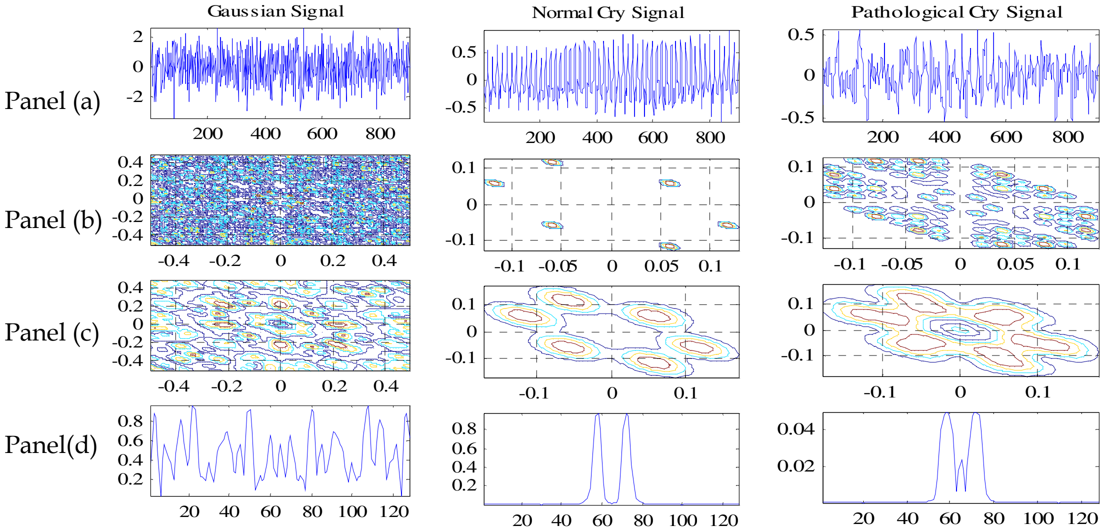


Figure 6.3. Bispectrum patterns for Gaussian signal, normal cry signal and pathological cry signal. In all the subfigures, Panel (a) time–domain waveforms, Panel (b) bispectrum using direct method, Panel (c) bispectrum using indirect method and Panel (d) diagonal slice derived from indirect method are shown for Gaussian signal, normal cry signal and pathological cry signal. In subfigures of Panel (a) X-axis is samples and Y-axis is amplitude, in subfigures of Panel (b) and Panel (c), X and Y-axis represents frequencies, ω_1 and ω_2, respectively and in subfigures Panel (d) X-axis is frequency and Y-axis is the amplitude of bispectrum.

6.3 Bispectrum Estimation

Estimation of bispectrum from the data is performed using two conventional methods, namely, (1) direct method and (2) indirect method. Description of these two methods is given below [**167**].

6.3.1 Indirect Method

Let the given dataset is $S(1)$, $S(2)$,….,$S(k)$. To estimate the bispectrum, following steps are followed:

1. Segment the infant cry data of length N into K segments of M samples each, *i.e.*, $N=KM$. Let these segments are denoted as $x(1)$, $x(2)$,…, $x(K)$.

2. Obtain *3rd* order estimate of moment for each segment after subtraction of its mean value as shown below

$$r^i(m,n) = \frac{1}{M}\sum_{l=s_1}^{s_2} x^i(l)x^i(l+n), \tag{6.6}$$

   where $i = 1,2,...,K$, $s_1 = \max(0,-m,-n)$ and $s_2 = \min(M-1, M-1-m, M-1-n)$.

3. Average the moment function over all *K* segments.

$$c_3^x(m,n) = \frac{1}{K}\sum_{i=1}^{K} r^i(m,n), \tag{6.7}$$

4. Generate the bispectrum estimate, *i.e.*,

$$B_3^x(\omega_1,\omega_2) = \sum_{m=-L}^{L}\sum_{n=-L}^{L} c_3^x(m,n)W(m,n)\exp(-j(\omega_1 m + \omega_2 n)), \tag{6.8}$$

   where $L<M-1$ and $W(m,n)$ is a two-dimensional (*i.e., 2-D*) window function.

6.3.2 Direct Method

1. Segment the infant cry data of length *N* into *K* frames of length *M* as mentioned in the indirect method using step *1*, add zeros at the end of each segment to make its length convenient for FFT computation, such that $M=2^l$, $l \in Z^+$.
2. Take DFT of each of the *K* segment, *i.e.*,

$$X^i(\lambda) = FFT(x^i(k)), \tag{6.9}$$

3. From the DFT coefficients, estimate of bispectrum of each segment using

$$b^i(\lambda_1,\lambda_2) = X^i(\lambda_1)X^i(\lambda_2)X^i(\lambda_1+\lambda_2), \tag{6.10}$$

4. Estimate of the bispectrum of the given data is the average of the bispectrum estimates given as

$$B_3^x(\omega_1,\omega_2)=\frac{1}{K}\sum_{i=1}^{K}b^i(\omega_1,\omega_2)\,. \tag{6.11}$$

Here, $\omega=(\frac{2\pi F_s}{N_0})\lambda$ and N_0 is the total number of samples in a segment.

Examples of bispectrum derived using direct and indirect methods are shown in Figure 6.3 for normal and pathological infant cry. The difference in the nature of bispectrum of normal and pathological infant cries can be observed. In particular, the spectral peak locations and the strength of peaks are different in normal and pathological infant cries. In normal infant cries, the strength of bispectrum peaks is higher and the bispectrum pattern is smooth. On the other hand, pathological cry samples show a higher number of spectral peaks in bispectrum with lower strength. These differences give the motivation to see the effectiveness of the bispectrum-based features for classification of normal *vs.* pathological infant cry.

In our experiments, we have compared the performance of bispectrum features derived from the direct and indirect methods with conventional feature extraction methods, *i.e.*, spectral features indicating implicit vocal tract system information.

6.4 Feature Extraction from Bispectrum

In this Section, three methods to extract features from bispectrum are presented.

6.4.1 Method A- Using Triangular Symmetry of Bispectrum

In [**168**], the triangular symmetry of bispectrum is used and features in the triangular region of the first quadrant are calculated. The features are defined as follows:

$$E_1^j = \{\sum_i |B(\omega_1^j,\omega_2^i)|\} \text{ where } i = \begin{cases} \frac{N}{2} - j + 1, \ldots, \frac{N}{2}, if\ \ j = 1, \ldots, \frac{N}{4}; \\ j, \ldots, \frac{N}{2},\ if\ j = \frac{N}{4} + 1, \ldots, \frac{N}{2}. \end{cases} \quad \text{(6.12}$$

$$\{E_2^{i-N/4}\} = \left\{\sum_j |B(\omega_1^j,\omega_2^i)|\right\}, \text{ where } j = \frac{N}{2} - i + 1, \ldots, i,\ \ if\ i = \frac{N}{4} + 1, \ldots, \frac{N}{2} \quad (6.13)$$

$$\text{and } E = \left\{ E_1^j, E_2^{i-\frac{N}{4}} \right\}. \quad (6.14)$$

In both the equations, *N* is the number of points in bispectrum. The feature E_1^j represents a feature vector formed by the summation of magnitudes of bispectrum at every column of the triangular region. The feature $E_2^{i-N/4}$ is a feature vector formed by the summation of magnitudes of the bispectrum in the triangular region. The two features combined are taken as an effective feature for classification task by the authors in [**168**]. This feature set captures peaks in both the frequency dimensions. However, it gives very large dimension feature vectors (in our case *1* x *256*).

6.4.2 Method B- Peaks, Peak Locations and Entropy

In this method, the first diagonal slice is obtained. On this diagonal slice, peak amplitudes and their locations are recorded as $\{\alpha_i, \omega_i\}$ where *i=1,2,3*. Top three peaks (a_i) and their corresponding locations ω_i are recorded. The normalized entropy of the bispectrum is defined as [**169**]:

$$P_1 = -\sum_n p_n \log p_n, \quad (6.15)$$

where $p_n = |B_x(\omega_1,\omega_2)| / \sum_\Omega |B_x(\omega_1,\omega_2)|$ and normalized bispectrum squared entropy is defined as $P_2 = -\sum_i p_i \log p_i$, where $p_i = |B_x(\omega_1,\omega_2)|^2 / \sum_\Omega |B_x(\omega_1,\omega_2)|^2$ distance between two peaks is defined as *d*, impulse width at *-3 dB* is denoted by *w* and energy of slice spectrum *E* is defined as $E = \sum_{i=1}^{M} |B_x(\omega_1,\omega_2)|^2_{\omega_1 = c\omega_2}$. The feature vector was defined by authors as [**169**]:

$$F = \{a_1, \omega_1, a_2, \omega_2, a_3, \omega_3, d, w, E, P_1, P_2\}. \tag{6.16}$$

Here, authors have considered two frequencies corresponding to a peak. However, as can be seen from bispectrum shown in Figure 6.3, both frequencies will have same value due to symmetry property. Hence, we have considered only one frequency component in order to remove redundancy in feature representation in feature space. This method has the advantage of having small computation time due to smaller dimension of feature vector.

6.4.3 Method C- Diagonal Slice of Bispectrum

In most of the bispectrum applications, a diagonal slice of bispectrum is used as a feature vector [17]. Diagonal slice is defined as bispectrum calculated on points where $\omega_1 = \omega_2$, *i.e.,*

$$D(\omega) = | B(\omega_1, \omega_2) |_{\omega_1 = \omega_2} . \tag{6.17}$$

Diagonal slice of bispectrum are shown in Figure 6.3 (d). This feature reduces the computational load at the cost of losing useful information in bifrequency planes. In our experiments, to extract the useful features from the bispectrum of the signals, higher-order singular value decomposition method is used. The higher-order singular value decomposition (HOSVD) algorithm is explained in brief here.

6.4.4 Higher-Order Singular Value Decomposition (HOSVD)

Here, HOSVD is proposed for feature extraction from bispectrum of infant cry signals. The higher-order singular value decomposition (HOSVD) theorem proposed in [**170**] is used to reduce the dimensionality of the feature space. This feature extraction method has recently been used by the authors for phoneme classification and normal *vs.* pathological cry classification [**171**], [**172**], [**173**], [**174**] . HOSVD is a generalization of SVD applied to a tensor. Initially, all the features are stacked together one after another to form a *3-D* tensor *A*. The *2-D* feature set, *i.e.*, bispectrum is in the space $F \in R^{I_1 \times I_2}$ and let the number of samples be I_s. The tensor A (of dimension $I_1 \times I_2 \times I_s$) can be represented in HOSVD form (as shown in Figure 6.4) by using eq. (6.18):

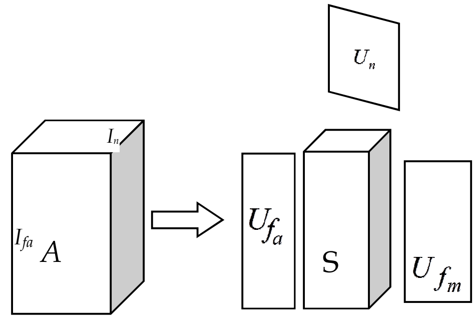


Figure 6.4. Singular value decomposition of tensor A . After [**170**].

$$A = S \times_1 U_{I_1} \times_2 U_{I_2} \times_3 U_S, \tag{6.18}$$

where S is the core tensor with the same dimension as A.

$U_{I_1} \in R^{I_1 \times I_1}$, $U_{I_2} \in R^{I_2 \times I_2}$ and $U_{I_s} \in R^{I_s \times I_s}$ are the unitary matrices of the corresponding subspaces of I_1, I_2 and I_s. The matrices U_{I_1} and U_{I_2} contains n mode singular vectors, *i.e.*,

$$U^{(n)} = [U_1^{(n)} \quad U_2^{(n)} \quad .... \quad U_{I_n}^{(n)}]. \tag{6.19}$$

The matrices U_{I_1} and U_{I_2} can be obtained from the matrix unfolding of A. The unfolded matrices $A_1 \in R^{I_1 \times I_2 I_S}$ and $A_2 \in R^{I_2 \times I_1 I_S}$ are obtained (as shown in Figure 6.5) and they are decomposed in their SVD representations to give U_{I_1} and U_{I_2} . Only first R_1 and R_2 principal components are retained from these unitary matrices, respectively. Next, $\hat{U}_{I_1} \in R^{I_1 \times R_1}$ and $\hat{U}_{I_2} \in R^{I_2 \times R_2}$ are obtained, which gives reduced dimension feature set, namely,

$$Z = B \times_1 \hat{U}_{I_1}^T \times_2 \hat{U}_{I_2}^T = \hat{U}_{I_1}^T.B.\hat{U}_{I_2}, \tag{6.20}$$

where $Z \in R^{R_1 \times R_2}$ and $B \in R^{I_1 \times I_2}$ which is taken from A. In order to classify infant cries as normal *vs*. pathological cries, the cry recordings are divided in appropriate cryunits. A cryunit is defined as a cry sound produced by an infant during the expiratory phase. The cryunits for an infant are similar to the words spoken by a speaker in speech recognition or other speech processing tasks.

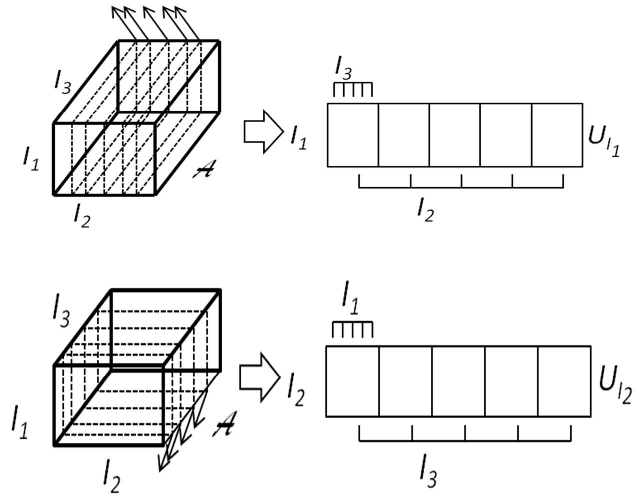


Figure 6.5. Unfolding of tensor A to matrix A_1 and matrix A_2. After [**170**].

6.5 Performance Measures

In the experiments, reported in Section 6.6- Section 6.8, following performance measures are used:

a) **Average classification accuracy**: The average classification accuracy (in %) is defined as the ratio of the correct classified samples to the total number of samples multiplied by *100*. Higher is the classification accuracy, better is the system's performance.

b) **Standard deviation of classification accuracy (σ)**: Standard deviation of the accuracy is defined as the square root of the variance. This should be as low as possible (ideally *0*).

c) ***95* % confidence interval**: Band of confidence for *95* % confidence interval (B) is ,

$$B = 1.96\sqrt{\frac{p(100-p)}{n}}, \tag{6.21}$$

where p is the identification rate or accuracy (in %) and n is the total number of genuine trials performed in order to obtain identification rate, *i.e.*, p and confidence interval is $[p - B,\ p + B]$. Here, *1.96* is used because for normal distribution with mean *0* and variance *1*, the area under the curve between *-1.96* and *+1.96* is *0.95* (*95* % of total area) [**149**]. That is one of the properties of probability density function (*pdf*). Confidence interval should also be as small as possible.

d) **Matthews's correlation coefficient (MCC)**: It is a measure of binary classification. This coefficient was introduced by a bio-chemist Brian W. Matthews in *1975*. It is defined as follows [**175**]:

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(FN+TP)(FP+TN)(FP+TP)(FN+TN)}}, \tag{6.22}$$

where *TP*= true positive, *TN*= true negative, *FP*= false positive and *FN*= false negative. The values of MCC lie between *-1* to *+1*. The value close to *+1* indicates the best classification performance and the MCC score *0* indicate better than random prediction and the value of MCC near to *-1* indicate totally incorrect classification.

e) **Probability excess (PE)**: PE is defined as

$$PE = \frac{TP \times TN - FN \times FP}{(FN+TP) \times (FP+TN)}. \tag{6.23}$$

The probability excess also varies between *0* (for random picking) and *1* (perfect prediction). In addition, *PE* is independent of the relative class frequency in the test set. Indeed, the probability excess can be shown to be simply *PE* = sensitivity + specificity – *1* [**176**].

f) **Specificity (SP)**: This is the percentage of the negatives classified as negative. In our case, it corresponds to correct detection of pathological cases. It is given by

$$SP = \frac{TN}{TN + FP}. \tag{6.24}$$

g) **Sensitivity (SN)**: This indicates the proportion of positive class classified as positives. This is equivalent to classifying healthy infants classified as healthy. SN is defined as $SN = \frac{TP}{TP + FN}$. PE, SP and SN should be close to one for a good classifier.

6.6 Classification of Normal and Pathological Infant Cries

6.6.1 Database

In this experiment, *Corpus I* is used [**91**]. Cry types are considered in two classes, namely, 1. normal and 2. pathological infants (suffering from any disease).

6.6.2 Experiment -1

6.6.2.1 *Experimental Setup and Feature Extraction*

This database is then divided into train and test dataset with a train-test ratio of *75-25*. Four such train-test datasets were created. From the cryunit signal, voiced segments are extracted using an energy-based algorithm (l^2 energy). The voiced data is segmented into non-overlapping frames of length *10* ms each. Each frame is normalized by subtracting the mean then bispectrum is calculated. From each of the infant cry signal, maximum *20* frames are taken.

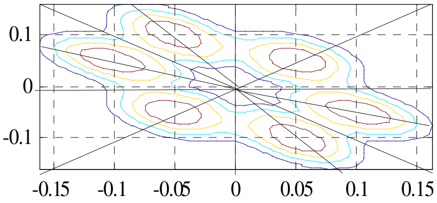


Figure 6.6. Symmetry regions of bispectrum.

Bispectrum is a two-dimensional (*i.e.*, *2-D*) feature of size *512* x *512* (for number of Fourier transform points of *128*) as shown in Figure 6.6. It can be observed from Figure 6.6, that bispectrum plot has twelve symmetry regions. The symmetry property of bispectrum follows from the properties of moments [**167**]. From Figure 6.6., it can be observed that the information of bispectrum in either first quadrant or in the third quadrant is sufficient to consider it as a feature set. In addition, this saves the memory space and reduces the computational complexity and time for feature extraction. In our experiments, we have considered information contained in third quadrant only. It reduces the feature size from *512* x *512* to *128* x *128* and then a tensor is formed for all frames of both *train* and *test* datasets. On these tensors, HOSVD is applied which reduces the size of bispectrum from *128* x *128* to *10* x *10*. These features are then stored as the logarithm of bispectrum feature vectors (*1* x *100*) for classification purpose. On the reduced feature set (*i.e., 1* x *100*), entropy is used as feature selection criterion. Using this criterion, feature size is reduced to [*3 5 8 10 20 30 40 50 60 70 80 90 100*]. On the reduced feature set, classification accuracy is determined using support vector machine (SVM) classifier with radial basis function (RBF) kernel. LIBSVM tool is used in our work [**177**]. The details of SVM are given in Appendix D.

6.6.2.2 Experimental Results

Classification performance of bispectrum features with different feature vector sizes is shown in Table 6.1. From Table 6.1, it can be observed that as the feature size increases, classification accuracy increases. However, after feature vector size of *40,* accuracy decreases. This trend is the observed because with increasing feature size, redundancy in features also increases which results in performance degradation. Another effect of increasing feature size is that as feature size increases, more pathological samples are recognized as normal samples by the classifier, *i.e.*, the distinction in feature vectors of two classes becomes ambiguous.

Confusion matrix for classification of normal and pathological infant cries using the proposed bispectrum- based features (for feature vector of size *1* x *30*) is shown in Table 6.2. It can be observed that *99.53* % frames of normal infant cries and *98.43* % frames of pathological infant cries from test dataset are classified correctly.

Table 6.1. Classification accuracy (in %) for *4*-fold cross-validation using holdout method

Feature size		3	5	8	10	20	30	40	50	60	70	80	90	100
Test set	Test_1	78.88	90.46	92.81	94.72	97.07	99.12	99.56	98.68	98.24	97.654	96.77	95.30	94.13
	Test_2	84.36	96.20	97.95	98.83	99.85	100.00	99.56	99.27	99.12	98.83	98.10	97.08	96.64
	Test_3	78.62	95.02	97.22	98.24	98.98	99.41	99.56	99.41	98.54	98.39	97.80	97.80	97.66
	Test_4	78.65	91.96	94.01	95.91	97.08	97.22	97.08	97.08	97.08	96.64	95.91	94.44	92.98
	Mean	**80.13**	**93.41**	**95.50**	**96.92**	**98.24**	**98.94**	**98.94**	**98.61**	**98.24**	**97.88**	**97.14**	**96.16**	**95.35**

*****In all experiments, one of the group (Test_1, Test_2, Test_3, Test_4) is taken for testing and remaining for training the classifier.

Table 6.2. Confusion matrix of classification of normal and pathological cries using bispectrum as a feature set. Adapted from [**171**]

Identified as		Normal	Pathological
Actual	Normal	425	2
	Pathological	4	251

Table 6.3. Classification accuracy (in %) with MFCC, LPC, PLP and bispectrum features. Adapted from [**171**]

Feature	Classification Accuracy
MFCC	*53.99*
LPC	*63.07*
PLP	*63.14*
Bispectrum	***98.94***

A comparison of the performance of proposed features with the state-of-the-art methods, namely, Mel frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC) and perceptual linear prediction coefficients (PLP) is shown in Table 6.3. Under the same experimental setup, classification accuracy is found. From Table 6.3, we can infer that classification with proposed feature outperforms with baseline features. MFCC gives a classification accuracy of *53.99* %, LPC and PLP gives average classification accuracy of *63.07* % and *63.14* %, respectively, whereas bispectrum gives a classification accuracy of *98.94* %. Bispectrum captures nonlinearity in the signal as a feature and hence, it performs better. The HOSVD theorem retains the principal components of the bispectrum. Since the principal components are used as feature vector, performance is better than other existing methods.

To quote statistical significance of our experimental results, *95* % confidence interval is also reported [22]. Figure 6.7 shows the plot of the confidence interval for different feature sizes. The confidence interval at a feature size of *1* x *30* is ± 0*.7692 %* around classification accuracy of *98.94 %* which indicates *better* performance and better statistical confidence of the proposed features compared to state-of-the-art methods.

6.6.2.3 Summary and Conclusions

In this work, it was found that features derived from bispectrum can classify normal and pathological infant cries better than the conventional state-of-the-art spectral features. The motivation behind using bispectrum is to use a feature which can capture nonlinearity in cry production mechanism as cry is a non-stationary and infant cry production system is nonlinear. In addition, bispectrum has the ability to detect noise in cry signal. In the case of pathological cries, the amount of noise is even higher (which is also apparent from use of *jitter* and *shimmer* for pathological signal classification). Increasing the number of cry samples in the pathological database may increase the classification accuracy (in %) for higher feature dimensions.

6.6.3 Experiment-2

6.6.3.1 Experimental Setup

We have *957* and *318* feature vectors of *45* and *16* infants, respectively, in the training and testing dataset of normal infant cries of *Corpus I*. In the pathological cries, we have *581* and *193* feature vectors of *28* and *10* infants for train and test datasets, respectively. Classification accuracy with different feature sizes has been determined and best classification accuracy is observed at feature vector size of *1*x*90*, which is *70.65 %*. Similar performance is observed for the proposed feature for a radial basis function (RBF) kernel with *γ=0.01* (classification accuracy of *70.86 %).*

6.6.3.2 *Experimental Results*

Table 6.4. Classification accuracy (in %) with MFCC, LPC, PLP and bispectrum features

Feature	Classification Accuracy (in %)
MFCC	52.41
LPC	61.27
PLP	57.41
Bispectrum	70.65

From Table 6.4, we can infer that classification with proposed feature *outperforms* with baseline features. MFCC gives a classification accuracy of *52.41 %*, LPC and PLP gives average classification accuracy of *61.27 %* and *57.41 %*, respectively, whereas bispectrum gives a classification accuracy of *70.65 %*. Bispectrum captures nonlinearity in the signal as a feature and hence, it performs better. The HOSVD theorem retains the *principal* components of the bispectrum. Since the principal components are used as a feature vector, performance is better than other existing methods. Confusion matrix for classification of normal and pathological infant cries using the proposed bispectrum- based features (for feature size *1*x *90*) is shown in Table 6.5. It can be observed that *72.64 %* frames of normal infant cries and *65.66 %* frames of pathological infant cries from test dataset are classified correctly.

Table 6.5. Confusion matrix of classification of normal and pathological cries using bispectrum as a feature

Identified as		Normal	Pathological
Actual	Normal	231	87
	Pathological	63	130

Table 6.6. Comparison of classification performances (in %) of bispectrum features extracted from existing methods under same experimental setup. Adapted from **[172]**

Feature	Feature Size	Average Accuracy (%)	*95* % Confidence Interval (%)
Method *A*	1x512	59.295	±4.23
Method *B*	1x512	60.425	±4.21
Method *C*	1x11	60.425	±4.21
Bispectrum	1x90	**70.65**	**±3.91**

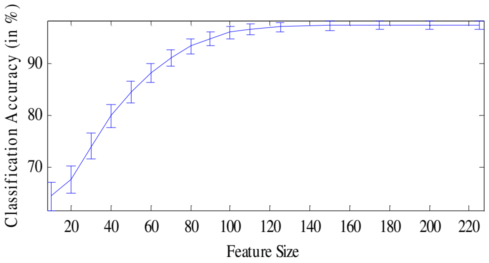


Figure 6.7. Classification accuracy (in %) of bispectrum-based features for different feature sizes with *95 %* confidence interval shown with bars.

Table 6.6 shows the classification performances of existing feature extraction methods with proposed method. Results show that proposed method of feature extraction using HOSVD gives excellent performance compared to existing methods along with a comparatively small feature dimension. Comparison with *method B* and *method C*, indicates that diagonal slice or features derived from it are not of sufficient importance in the classification of normal and pathological infant cries. To quote *statistical significance* of our experimental results, *95 %* confidence interval is also reported in Figure 6.7. The confidence interval at a feature size of *1* x *90* is ± *3.91 %* around classification accuracy of *70.65 %* which indicates better performance of the proposed features compared to state-of-the-art methods.

6.6.4 Experiment -3

6.6.4.1 Segmentation of Cry Utterance (episode) into Cryunits

Energy-based (*i.e.*, l^2 norm) algorithm is applied to the cry signal of an infant, to automatically extract the cryunits. The algorithm is proposed in a study reported in [**52**] and has been used by researchers for segmentation of infant cry in small cryunits. For this purpose, the cry signals are first downsampled from *44.1* kHz to *18* kHz. The cry signal is then passed through a lowpass filter with cutoff frequency of *6* kHz to remove the possible noise added by the microphone during recordings of infant cries. On this filtered signal, amplitude normalization is performed to reduce the variance among cryunits and hence, among features derived from cry signals. This pre-processed signal is then divided into cryunits. To extract useful cryunits from a cry utterance, cry signal is segmented into frames of *10* ms with *50* % overlap. For each of the frame, short-time Fourier transform (STFT) is computed and energy is computed as shown in Figure 6.8. The STFT is a representation of the signal in time-frequency planes. It is defined as:

$$X(k,\omega)=\sum_{n=-\infty}^{\infty} x[n]w[n-k]e^{-j\omega n}, \tag{6.25}$$

where $X(k,\omega)$ is the STFT of the signal $x[n]$ and $w[n]$ is the window function. The parameters k and ω represents the time and frequency, respectively. In our experiments, a rectangular window is used for STFT computation because energy is computed in time-domain and windows other than rectangular window add tapering effect to the signal under consideration in time-domain.

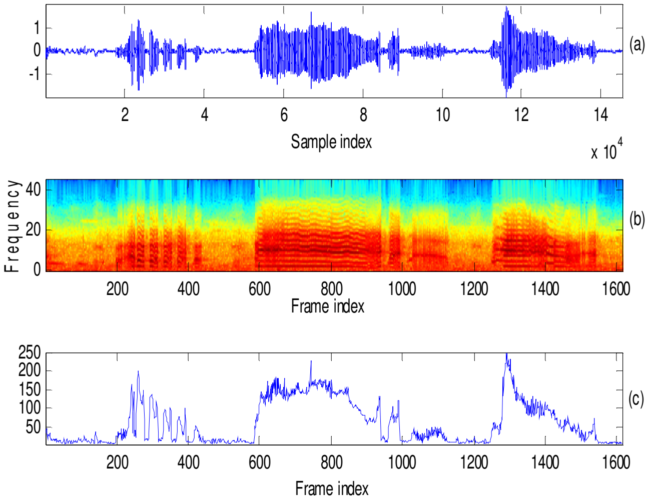


Figure 6.8. (a) Time-domain waveform of infant cry, (b) corresponding narrowband spectrogram, (c) short-time energy of the signal.

The study reported in [**52**] proposed *25* % of the maximum energy to be taken as a threshold. For our database, this threshold reduces the number of cries significantly because we have considered duration parameter as well in order to remove infant cry produced due to inspiration and hiccups. If the energy of a cry frame is more than the threshold, then the cry segment is considered, otherwise it is ignored. In this manner, if the number of frames contributes to a cryunit which is more than or equal to *0.75* sec. long (because in *0-1* year old infants, respiration rate is *30-60* sec.), then that particular cryunit is stored else it is omitted from the experiment (as shown in Figure 6.9).

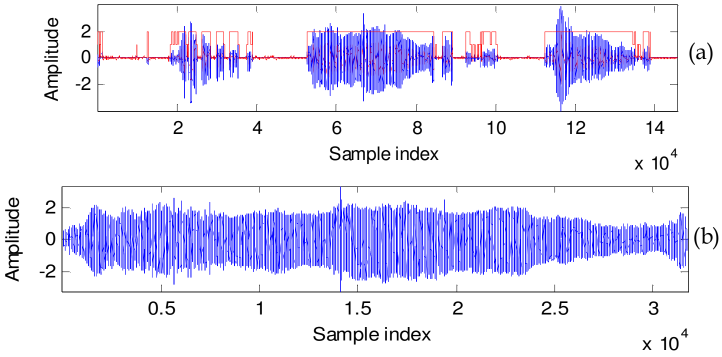


Figure 6.9. (a) Time-domain waveform and short-time energy of infant cry, (b) cryunit extracted from the signal shown in (a).

This procedure is followed on both the classes, *i.e.,* normal infant cries *vs.* pathological infant cries. Because of variability in age, weights and health condition in infant group under study, durational feature set does not carry significance. After getting cryunits from the cry recordings, feature extraction from bispectrum and classification is performed as illustrated in the next Section.

6.6.4.2 *Experimental Setup*

This database is then divided into *train* and *test* dataset with a ratio of *75:25.* The cry signal is first converted into cryunits. For the experiments reported in this experiment, the cryunits which are longer than *0.75* sec. are considered. Small cryunits are not considered because these may include sounds which are produced during inhalation and hiccups.

The cryunit is segmented into frames of length *50* ms each. For each cry segment, bispectrum is calculated using HOSA toolbox available online **[178]**. Bispectrum is a two-dimensional (*i.e., 2-D*) feature of size *256* x *256* as shown in. It can be observed from that bispectrum plot has twelve symmetry regions.

The symmetry property of bispectrum follows from the properties of moments. The third-order cumulant has six symmetry regions, *i.e.*, $c_{3,x(\tau_1,\tau_2)}$ = $c_{3,x(\tau_2,\tau_1)}$ = $c_{3,x(-\tau_2,\tau_1-\tau_2)}$ = $c_{3,x(-\tau_1,\tau_2-\tau_1)}$ = $c_{3,x(\tau_2-\tau_1,-\tau_1)}$ = $c_{3,x(\tau_1-\tau_2,-\tau_2)}$. From Figure 6.3, it can be observed that the information of bispectrum in either first quadrant or in the third quadrant is sufficient to consider it as a feature set. This saves the memory space and reduces the computational complexity and hence, the feature extraction time as well (as discussed in Section 6.4.4). In our experiments, we have considered information contained in first quadrant only. It reduces the feature size from *256* × *256* to *128* × *128* (in the case of direct bispectrum computation method, where FFT size is taken as *128*) and from *128* × *128* to *64* × *64* (in the case of the indirect method of bispectrum estimation). Then, a tensor (per infant cryunit) is formed for both train and test datasets. For each of the cryunit, first *25* frames are considered to form a tensor in order to maintain the same contribution from each of the cryunit because of their varying durations. After reduction using HOSVD to *10* × *10,* a feature vector is formed of the dimension *1* × *100.* After this, DCT is applied to the obtained bispectrum features.

For the classification task, LIBSVM tool is used in this work [21]. In the experiments, infant-independent *4*-fold cross-validation is performed. For the classification task, radial basis function (RBF) kernel is used in support vector machine (SVM) classifier. The parameters selected for the kernel are used as standard setup parameters given in LIBSVM. All the results reported in this Section are independent of earlier results quoted in experiment 1 and experiment 2 [**172**], [**171**]. In particular, here cryunits are formed automatically. However, in cited papers [**171**], [**172**], cryunits were not formed and preprocessing was not done which resulted in mean classification accuracy of *70.65* % with ± *3.91* confidence interval (*95* %).

6.6.4.3 Experimental Results

Experimental results are shown in Table 6.7 to Table 6.24. All quoted results show the classification performance after *4*-fold cross-validation. For cross-validation, infant cry signals are divided into four equal groups (infant-independent) and out of these groups, one group is kept for testing and remaining three groups are used for training. The number of feature vectors may not be the same in all the groups because the number of cryunits generated from the cry utterances is not equal (possible reason for this could be either the length of cries is not equal and/or the cryunit criterion may not be satisfied). The number of infants and the cryunits in each group are as follows: in the normal infant cries group *1* has *16* infants and *80* cryunits, group *2* has *16* infants and *62* cryunits, group *3* has *16* infants and *74* cryunits and group *4* has *14* infants and *48* infant cries. In the pathological class, each of the four groups has *10, 10, 10* and *9* infants and *45, 40, 30* and *34* cryunits, respectively. The cross-validation is done to check the performance of the system which is independent of the particular set of infants and to show the robustness of the features against infant-dependency because the data size is small.

6.6.4.4 Comparison with standard features

In the experiment, the performance of the bispectrum features is compared against state-of-the-art spectral feature sets, namely, MFCC, LPC, PLP, LFCC, LPCC and PLPCC. It has been found in earlier experiments reported in [**172**] and [**171**], that the feature dimension of 1×100 is good to give optimum results. Hence, in our experiments, we have considered bispectrum dimension reduction to 10×10 using HOSVD and then using 1×100 feature vector for the present classification task.

Table 6.7. Comparison of classification accuracy (in %) using different feature sets for the classification of normal and pathological infant cries

Feature Set	Dimension of Feature vector	Classification Accuracy (in %)	std.
Bispectrum (Direct Method)	1 x 100	**82.44**	4.03
Bispectrum (Indirect Method)	1 x 100	**81.65**	4.28
LPC	1 x 19	63.01	5.22
PLP	1 x 9	60.98	5.96
LFCC	1 x 12	51.79	3.28
MFCC	1 x 12	48.52	7.26
LPCC	1 x 19	58.86	1.49
PLPCC	1 x 9	63.05	5.22

std.= Standard deviation

Table 6.7 shows the classification performance of bispectrum and conventional feature extraction methods. It can be observed from Table 6.7 that the traditional feature extraction methods such as LFCC, MFCC, LPC, LPCC, PLP and PLPCC performs poorly compared to the proposed bispectrum features for classification of normal and pathological infant cries. The highest classification accuracy achieved with bispectrum features is *82.44* % using the *direct* method of bispectrum computations. Because the difference in classification performance of the bispectrum estimation methods (*i.e.*, direct *vs.* indirect) is not very significant; for comparison of different feature extraction methods from bispectrum, the indirect method is used in remaining set of experiments presented in this thesis. The standard deviation (*std.*) in all the experiments is high because, in one of the groups, the classification performance was dropped significantly. The reason for the drop in classification accuracy is that, in that group, the background noise was very high which resulted in high standard deviation in results. Classification performance of the bispectrum features using indirect method varies with the lag and the Fast Fourier Transform (FFT) bin sizes. The variation of classification accuracy with lag and FFT bin sizes is shown in Table 6.8. It can be observed from Table 6.8 that the classification accuracy is highest with FFT size of *128* and lag *16*. However, lag *8* and FFT size of *128* is considered in all the experiments because the difference in the classification accuracy is not significant. However, it reduces the bispectrum size significantly which helps in reducing the computation time.

Table 6.8. Classification accuracy (in %) with different FFT bin sizes and lags in bispectrum estimation using the indirect method.

FFT size	Lag	Classification accuracy (in %)
256	16	81.61
128	16	**82.44**
128	8	81.27
128	4	36.97

Table 6.9. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using bispectrum (Direct method)

	Normal	Pathological
Normal	**83.66**	16.33
Pathological	20.12	**79.87**

Table 6.10. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using bispectrum (Indirect method)

	Normal	Pathological
Normal	**83.51**	16.48
Pathological	20.81	**79.18**

Table 6.11. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using MFCC feature set

	Normal	Pathological
Normal	73.04	26.95
Pathological	83.82	16.17

Table 6.12. Confusion matrix for classification of normal *vs.* pathological cry classification (in %) using LPC feature set

	Normal	Pathological
Normal	100	**0**
Pathological	100	**0**

Table 6.9 and Table 6.10 show confusion matrices of the bispectrum features derived using direct method and indirect methods, respectively. It can be observed that the classification accuracy of the bispectrum-based features derived using direct method is higher than features derived from bispectrum estimated using the indirect method. Comparison of classification performance with MFCC and LPC feature sets is shown in Table 6.11 and Table 6.12, respectively. Compared to pathological infant cries classification,

all the features considered here are found to be good in capturing characteristics features of normal infant cry. It can be observed that the MFCC features perform poor in capturing distinct features of normal *vs.* pathological infant cries (for example*, 16.17* % pathological samples are classified correctly by MFCC). The same is observed by the authors during playback experiments of normal *vs.* pathological infant cries. Since MFCC features mimic the human auditory model, it performs poor in pathological infant cry classification. LPC features are found to perform very poorly in the classification of pathological and normal infant cries compared to the MFCC feature set. These features cannot identify the differences in normal and pathological cries.

6.6.4.5 Comparison with other feature extraction methods used for bispectrum

In Table 6.13, comparison of various feature extraction methods of bispectrum is given. It can be observed that the proposed HOSVD method performs better than the other feature extraction methods. In particular, for *Method A* (triangular symmetry- based features), it may happen that the sum of the bispectrum features in x and y directions (eq. (6.12) and eq. (*6.13*)) in normal and pathological cases comes to nearby close values. Hence, this feature set performs poor in normal and pathological infant cry classification work. Another drawback of this method is that dimension of the feature vector is comparatively higher than other methods. In *Method B*, features are extracted using peaks and peak locations, entropy and energy of the bispectrum as mentioned earlier in Section 6.4.2. The performance of these features is much better than the Method *A*. The advantage of this method is that it is fast in computation as the feature dimension is small. However, the performance of the *Method B* is lesser than the HOSVD method. The computation time for the feature extraction methods taken by the Pentium ® dual core processor, operating at *2.3* GHz, with *3* GB RAM and *32*-bits operating system are as follows: *Method A* took *0.86 sec., Method B* requires *1.13 sec* and HOSVD took

1.47 sec for a cry segment of *50 msec*. In HOSVD theorem for dimension reduction, a tensor is decomposed in its eigenvalues and eigenvectors. For dimension reduction, only those eigenvalues which are high in rank are kept (eigenvalues are in decreasing order in matrix *S*). Then, from these chosen eigenvalues and eigenvectors, feature dimension is reduced using eq. (6.20*)*. These dimension reduced feature vectors keep information specific to a class (*i.e.*, normal *vs*. pathological) because the singular values are evaluated over the tensor. This may result in the correct classification of the infant cries in normal and pathological classes.

Table 6.13. Comparison of performance of feature extraction methods from bispectrum (direct method) on classification accuracy (in %)

Feature Set	Feature Dimension	Classification Accuracy (in %)	std.
Method A	1 x 256	61.63	3.79
Method B	1 x 11	62.44	5.02
HOSVD	1 x 100	**81.64**	4.28

Method A: using the triangular symmetry of bispectrum, Method B: Peaks, Location of Peaks and Entropy.

6.6.4.6 Comparison of various features on other performance measures

In general, getting large enough samples of infant cries is a challenging task. Furthermore, getting pathological infant cry samples in a large number compared to normal samples is extremely difficult. Hence, in the database used in this thesis, the number of normal and pathological cry samples is not the same. However, showing statistical significance of results on the small dataset is necessary and hence, taken care by using *4*- fold cross-validation experiments. In order to deal with the imbalanced dataset, other statistical classification measures are necessary. From Table 6.12, we can observe that using LPC feature set, none of the samples is classified as a pathological signal. However, the classification accuracy for the LPC feature set is *63.01 %*. This shows that in the present problem, classification accuracy does not give a complete description of the classification results. In our experiments, we are using other measures such as MCC, PE, SP and SN to compare the performance of various feature sets considered in this chapter.

Table 6.14. Comparison of performances of various feature sets using various statistical parameters

Feature Set	MCC	PE	SP	SN
Bispectrum (Direct Method)	0.62	0.63	0.79	0.83
Bispectrum (Indirect Method)	0.60	0.61	0.77	0.83
LPC	0	0	0	1
PLP	0	0	0	1
LFCC	-0.10	-0.09	0.17	0.73
MFCC	-0.126	-0.123	0.22	0.654
LPCC	-0.02	-0.01	0.01	0.88
PLPCC	0	0	1	0

MCC: Mathews' correlation coefficient, PE: probability excess, SP: specificity, SN: sensitivity.

Table 6.14 shows the performance of various state-of-the-art methods and bispectrum-based features based on parameters derived from the confusion matrices defined earlier in Section 6.5. In the present problem of classifying pathological infant cries from the normal ones, the higher classification accuracy of pathological cries is desired as compared to normal infant cries. Because in realistic scenarios, the cost associated with the infant detected as normal for a pathological infant is much higher than the cost associated with normal infant diagnosed as pathological. In addition, the penalty or cost depends on the severity of the disease. The specified parameters MCC, PE, SP and SN remove this bias of unbalanced datasets. For a feature to be good in classifying the binary data, all proposed parameters should have values close to unity.

From Table 6.14, it can be observed that bispectrum is the *only* feature set which gives MCC close to *1* for all the parameters. On the other hand, MCC and PE are close to zero for other spectral feature sets such as LPC, PLP, LPCC, MFCC and PLPCC. All these features are good at identifying the normal class, however, very poor in identifying samples from pathological class.

6.6.4.7 Robustness of the bispectrum features in noisy conditions

Let $s(n)$ be the clean speech or cry signal and $n(t)$ be the noise signal then noisy speech signal under additive noise environment is given by

$$s_{\eta}(t) = s(t) + n(t). \tag{6.26}$$

where $n(t)$ can be additive babble, car, white and HF channel noise. If the probability density function for the noise under consideration is Gaussian, then when we take its bispectrum, it will be zero. It means, noise having Gaussian distribution get suppressed in bispectrum-domain and thus, bispectrum of noisy speech or noisy infant cry signal is given by

$$B_{s_n}(\omega_1, \omega_2) \approx B_s(\omega_1, \omega_2) + B_n(\omega_1, \omega_2) . \tag{6.27}$$

For noise with Gaussian distribution $B_n(\omega_1, \omega_2) = 0$ and hence,

$$B_{s_n}(\omega_1, \omega_2) \approx B_s(\omega_1, \omega_2). \tag{6.28}$$

Thus, bispectrum features have noise suppression capability. It is for this reason that bispectrum features give superior performance than the other spectral features such as LPCC, MFCC, *etc.*, under signal degradation or noisy conditions.

To show the robustness of the proposed features in the presence of noise or signal degradation conditions (as in realistic hospital environment noise is always present), the experiment was conducted with additive babble noise, car noise, HF channel noise and white noise with various signal-to-noise-ratio (SNR) levels. These four different types of noise samples are taken from the NOISEX-2002 database [**179**]. The effect of various noises on bispectrum features is shown in Table 6.15. We can observe almost similar performance in the presence of noise in both the feature extraction methods of bispectrum. At SNR as low as *-10* dB, the performance is almost similar. Experimental results for signal degradation conditions show that the proposed bispectrum features are noise robust than the other spectral features. Compared to white noise and babble noise, performance degradation with increasing SNR is more in car noise and HF channel noise.

Table 6.15. Effect of different noises on classification performance (in %) when indirect method of bispectrum is used

SNR (in dB)	Babble Noise (direct method)	Babble Noise (Indirect method)	White Noise	Car Noise	HF Channel
Without noise	**82.44± 4.04**	**81.65± 4.28**	**81.65 ± 4.28**	**81.65 ± 4.28**	**81.65 ± 4.28**
20	82.29± 3.81	81.37± 4.31	81.82 ± 4.29	81.57 ± 3.84	81.44 ± 4.30
15	82.44 ± 3.84	81.39± 4.59	81.82 ± 4.05	81.43 ± 4.21	81.66 ± 3.87
10	82.08 ± 3.82	81.73± 4.57	81.92 ± 4.27	81.39 ± 4.56	81.74 ± 3.64
5	81.81 ± 3.60	81.65± 4.12	81.64 ± 3.94	81.73 ± 4.25	81.79 ± 3.62
0	82.41 ± 3.54	81.83± 4.40	81.75 ± 3.95	81.16 ± 3.67	81.11 ± 3.41
-5	81.84 ± 4.35	80.72± 3.55	81.83 ± 3.17	80.38 ± 3.86	81.68 ± 3.66
-10	81.47 ± 4.61	81.22± 3.58	81.67 ± 4.05	80.27 ± 3.35	79.98 ± 4.11

Table 6.16. Effect of babble noise on classification performance (in %) on standard features

SNR (in dB)								
Feature Set	**Clean**	**20**	**15**	**10**	**5**	**0**	**-5**	**-10**
Bispectrum	**81.65**	**81.37**	**81.39**	**81.73**	**81.65**	**81.83**	**80.72**	**81.22**
MFCC	48.53	48.42	48.57	48.81	49.01	49.53	50.44	50.83
LPC	63.01	63.01	63.02	63.02	63.02	63.02	63.02	63.02
PLP	60.98	61.06	61.13	61.45	62.83	62.96	62.98	63.02
LFCC	51.79	51.93	52.02	52.38	52.93	53.64	54.13	54.56
LPCC	58.86	58.88	58.94	59.16	59.25	58.89	58.89	59.11
PLPCC	63.02	63.02	63.02	63.02	63.02	63.02	63.02	63.02

Table 6.17. Effect of babble noise on classification performance (in MCC) on standard features

SNR (in dB)								
Feature Set	**Clean**	**20**	**15**	**10**	**5**	**0**	**-5**	**-10**
Bispectrum	**0.61**	**0.60**	**0.60**	**0.61**	**0.61**	**0.61**	**0.59**	**0.60**
MFCC	-0.13	-0.13	-0.13	-0.13	-0.12	-0.11	-0.09	-0.09
LPC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PLP	-0.06	-0.06	-0.06	-0.06	-0.03	-0.01	-0.01	0.00
LFCC	-0.11	-0.10	-0.10	-0.10	-0.08	-0.07	-0.07	-0.08
LPCC	-0.02	-0.02	-0.02	-0.02	-0.01	-0.02	-0.01	-0.02
PLPCC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 6.18. Effect of car noise on classification performance (in %) on standard features

SNR (in dB)								
Feature Set	Clean	20	15	10	5	0	-5	-10
Bispectrum	81.65	81.57	81.43	81.39	81.73	81.16	81.38	80.27
MFCC	48.53	48.74	48.72	48.65	48.81	49.08	49.34	50.01
LPC	63.01	63.00	63.01	63.02	63.02	63.02	63.02	63.02
PLP	60.98	61.06	61.17	61.93	62.90	62.94	62.96	62.94
LFCC	51.79	51.78	51.88	52.11	51.87	52.31	52.56	52.86
LPCC	58.86	58.82	58.88	58.46	58.75	58.78	58.59	58.97
PLPCC	63.02	63.02	63.02	63.02	63.02	63.02	63.02	63.02

Table 6.19. Effect of car noise on classification performance (in MCC) on standard features

SNR (in dB)								
Feature Set	Clean	20	15	10	5	0	-5	-10
Bispectrum	**0.61**	**0.61**	**0.60**	**0.60**	**0.61**	**0.60**	**0.58**	**0.58**
MFCC	-0.13	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.11
LPC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PLP	-0.06	-0.06	-0.06	-0.05	-0.02	-0.01	-0.01	-0.01
LFCC	-0.11	-0.11	-0.10	-0.10	-0.10	-0.08	-0.07	-0.07
LPCC	-0.02	-0.03	-0.02	-0.04	-0.03	-0.03	-0.04	-0.03
PLPCC	0.00	0.00	0.00	0.00	-0.06	0.00	0.00	0.00

Table 6.20. Effect of HF Channel noise on classification performance (in %) on standard features

SNR								
Feature	clean	20	15	10	5	0	-5	-10
Bispectrum	81.65	81.44	81.66	81.74	81.79	81.11	81.68	79.98
MFCC	48.53	48.51	48.42	48.47	48.53	48.61	49.15	50.71
LPC	63.01	63.02	63.02	63.01	63.01	63.02	63.02	63.02
PLP	60.98	60.94	61.00	61.08	61.36	62.36	63.02	63.02
LFCC	51.79	51.86	51.99	51.95	51.81	52.11	52.37	52.87
LPCC	58.86	58.93	58.93	58.47	58.82	59.01	59.14	58.73
PLPCC	63.02	63.02	63.02	63.02	63.02	63.02	63.02	63.02

Table 6.21. Effect of HF Channel noise on classification performance (in MCC) on standard features

SNR								
Feature	clean	20	15	10	5	0	-5	-10
Bispectrum	**0.61**	**0.60**	**0.61**	**0.61**	**0.61**	**0.60**	**0.61**	**0.57**
MFCC	-0.13	-0.13	-0.13	-0.13	-0.14	-0.13	-0.13	-0.11
LPC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PLP	-0.06	-0.06	-0.06	-0.06	-0.05	-0.03	0.00	0.00
LFCC	-0.11	-0.10	-0.10	-0.11	-0.11	-0.11	-0.11	-0.12
LPCC	-0.02	-0.02	-0.02	-0.04	-0.03	-0.03	-0.01	-0.03
PLPCC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 6.22. Effect of white noise on classification performance (in %) on standard features

SNR								
Feature	Clean	20	15	10	5	0	-5	-10
Bispectrum	81.65	81.82	81.82	81.92	81.64	81.75	81.83	81.67
MFCC	48.53	48.54	48.48	48.46	48.43	48.42	48.38	48.36
LPC	63.01	63.02	63.02	63.02	63.02	63.00	63.00	63.00
PLP	60.98	60.90	60.90	60.93	61.00	61.12	61.34	62.12
LFCC	51.79	51.88	51.88	51.87	51.86	51.93	52.11	52.26
LPCC	58.86	58.70	58.77	58.89	59.03	58.82	58.73	58.49
PLPCC	63.02	63.02	63.02	63.02	63.02	63.02	63.02	63.02

Table 6.23. Effect of white noise on classification performance (in MCC) on standard features

	SNR							
Feature	Clean	20	15	10	5	0	-5	-10
Bispectrum	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**
MFCC	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.14
LPC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PLP	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.05	-0.05
LFCC	-0.11	-0.10	-0.10	-0.11	-0.11	-0.10	-0.10	-0.10
LPCC	-0.02	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.03
PLPCC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

For the other features as well, the effect of noise on classification performance is shown in Table 6.16 - Table 6.23 is shown using the classification accuracy (in % and MCC). It can be observed that performance for LPC and PLPCC features classification accuracy remains unchanged with SNR and noises. In these features, all samples are classified as normal resulting in MCC value of zero. In other features, classification accuracy increases with SNR in noisy environment except in the case ofHF channel noise. It can be observed from Figure 6.10 that the distribution of all the noises considered in this work has a Gaussian distribution. Hence, bispectrum features are performing well with these noises. With increasing SNR, more samples were classified as normal infant cry samples and the misclassification of pathological infant cry samples increased. This resulted in higher average classification accuracy (in %) though there is a decrease in correct classification of pathological infant cry samples. To show that the noises added are suppressed due to the Gaussian nature of the noise distribution, the histograms of the four noises are plotted in Figure 6.10. It is found that all these noises have a normal distribution, therefore, theses noises are suppressed by bispectrum features.

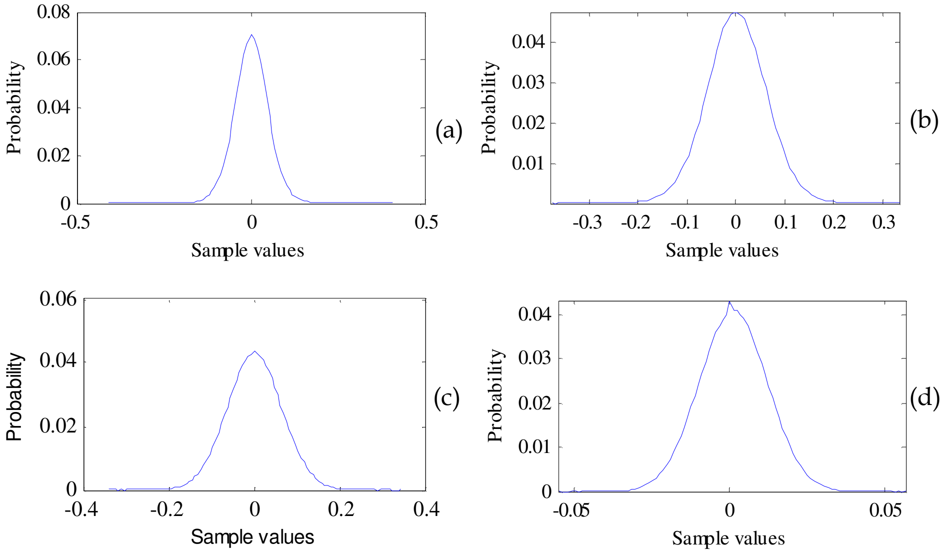


Figure 6.10. Probability density function (pdf) of additive noises. (a) babble, (b) car, (c) HF channel and (d) white noise.

6.6.4.8 *Effect of cryunit segmentation on classification performance*

To show the effectiveness of the cryunit segmentation, results reported in earlier work are shown in Table 6.24 [**172**] . It can be observed that the pre-processing of the signal and considering cryunits for feature extraction, results in tremendous improvement in classification performance (from *70.56* % to *82.44* %). The cryunits segmentation considers different moods (such as tiredness, increasing or decreasing pain, fussiness, *etc.*) and hence, different articulation of the vocal tract system which gives a complete details in the classification task.

Table 6.24. Classification accuracy (in %) with MFCC, LPC, PLP and bispectrum features. Adapted from [**172**]

Feature Set	Classification Accuracy (in %)
MFCC	52.41
LPC	61.27
PLP	57.41
Bispectrum	**70.65**

6.6.4.9 *Summary and Conclusions*

In this experiment, it is found that the proposed method of using HOSVD for feature extraction retains the useful information related to a specific class. In

HOSVD, the feature tensor is decomposed in its eigenvectors and to reduce the feature dimension, eigenvectors with higher eigenvalues are retained (initial *10 eigenvectors*) because they comprise most of the information contained in the tensor. Hence, it performs better than the other feature extraction methods. However, if the dimension of the feature vector is increased further, classification accuracy increases. It has been found that the reduced size of bispectrum (after application of HOSVD), gives a classification accuracy of *97.35 %*. In all the experiments, the standard deviation is high during four-fold cross-validation classification. For one of the groups, the classification performance dropped significantly resulting in high standard deviation. This may be because of improper selection of pathology in pathological class (a pathological condition which does not affect severely characteristics of infant cry) or higher background noise during data collection. Finally, proposed bispectrum-based features are found to have excellent noise robustness in noisy conditions of additive noise with various SNR levels.

After application of signal processing methods for the classification of normal and pathological infant cries, classification of pathological cries is attempted in the next Sections. In the next Sections of this chapter, pathologies considered are restricted to asthma and hypoxy ischemic encephalopathy (HIE). In the next Section, first these two pathological infant cries are classified from the normal infant cries and then the classification of asthma and HIE infant cries is also attempted.

6.7 Classification of Pathological Infant Cries

Asthma is a chronic *inflammatory* disease of the airways in which airways become blocked or narrowed. HIE is a condition in which brain does not receive enough supply of oxygen. Because of oxygen deficiency, brain cell may begin dying, resulting in brain damage from severe oxygen deficiency after the birth. Spectrographic analysis has been used in past for identification of HIE and asthma cry samples. However, it is not used for classification task [**94**].

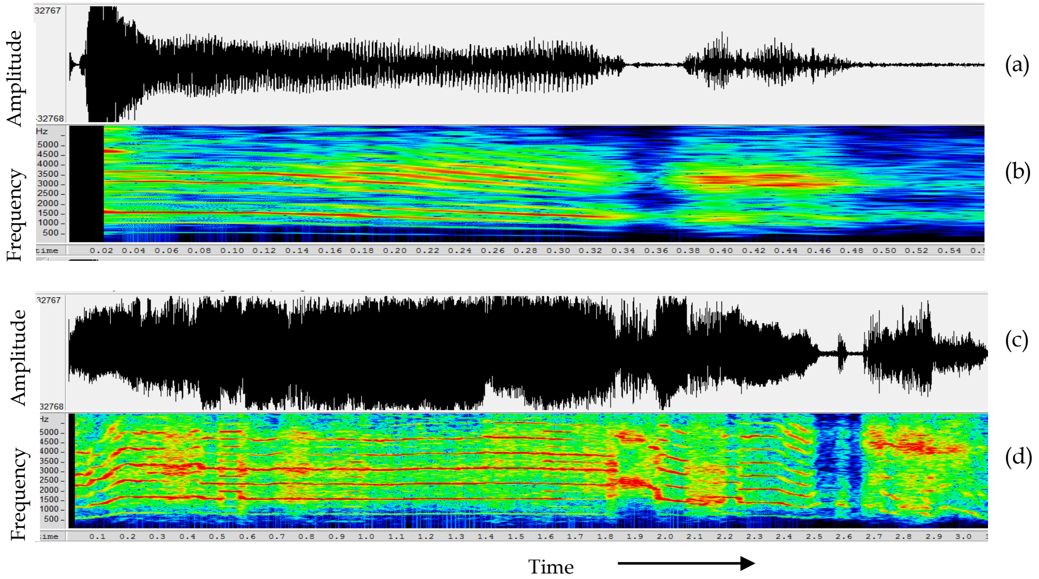


Figure 6.11. (a) Time-domain waveform, (b) corresponding spectrogram of asthma cry sample, (c) time-domain waveform and (d) corresponding spectrograms of HIE cry samples. In all the subfigures, X-axis represents time in sec.

Spectrograms of asthma and HIE cry samples are shown in Figure 6.11. Spectrograms of asthma shown in Figure 6.11 (b), it is similar to normal infant cry. Melody pattern is rising or falling. Inspiratory phonation is also present. In Figure 6.11 (d), the spectrogram of HIE cry is shown, which indicates the presence of *double harmonic breaks*. The duration of cry is also smaller than the asthma cry samples. The spectrographic analysis could not find its place in medical diagnosis because its accuracy depends on the expertise of the spectrogram reader. Thus, there is a need for a computer-based algorithm for pathology classification.

For the classification of HIE and asthma infant cries, features derived from four different approaches namely, modulation spectrogram, glottal inverse filtering, auditory spectrogram and modified group delay function are used. In addition, their relative performance comparison with state-of-the-art features is also shown in the thesis. In the experiments of cry classification, SVM classifier is used with RBF kernel function. RBF is selected because it is the most widely accepted kernel function for the classification task.

6.8 Classification of Normal, Asthma and HIE Infant Cries

In the following Sections of this chapter, attempts have been made to classify asthma infant cries from the HIE infant cries. It has been found that all the proposed features are performing well to classify these two pathologies from the cry signals. However, there is a need to develop an algorithm to classify these pathologies from the normal infant cries. Because this is the primary requirement of any of the infant cry analyzer developed for the purpose of giving the alarming sign for the presence of pathology in the infants. In the next few experiments, the classification task of normal infant cries, asthma infant cries and HIE infant cries is performed.

6.8.1 Database

Corpus II is used for this purpose. From there, all the normal cries irrespective of the reason of crying are considered in this class (in particular, *40* infant cries) and the distribution of asthma and HIE cries are same as those used in earlier experiments. All the infant cries are divided in the cryunits using the algorithm explained earlier. From the total cryunits, *75* % of the infant cries of infants in each group are taken for training and remaining for testing.

For the classification task, hidden Markov Model (HMM) classifier is used. In the HMM classifier, *5* states are considered, in each state, *3* Gaussian models are considered. The features used are MFCC, modulation spectrogram-based features, PLPCC and bispectrum-based features. Results of the classification are given in next Section.

6.8.2 Feature Extraction

Cryunits are divided into smaller duration frames (duration varying according to a feature set). For each of the cry signal frame, features are extracted. The specifications of the features used are as follows:

a. **Bispectrum**: The frame duration is taken as *100* ms with an overlap of *30 ms*. The indirect method of bispectrum estimation is used with a lag of *8* samples. The dimension of the estimated bispectrum is *128* × *128.* From this bispectrum, the diagonal slice is taken as the feature vector for classification of the three classes. The dimension of the feature vector is *1* × *128.*

b. **MFCC**: Standard MFCC feature vector with delta and delta-delta MFCC are considered. The 0^{th} coefficient of MFCC is also considered in the experiments, giving a feature vector of dimension *1* × *39* for a cry frame of *30* ms with *50* % overlap.

c. **PLPCC**: Standard PLPCC feature with *12* coefficients is considered in feature extraction process. The frame duration is taken as *30 ms* with an overlap of *50* %.

The performance is measured in % classification accuracy which represents the percentage of correctly classified samples from the total number of samples.

6.8.3 Experimental Results

The results of classification performances of the four features used are given in Table 6.25. The experimental results show that the performance of PLPCC features is comparatively better in classifying the three classes. Since the purpose of this experiment is to classify pathological cry samples from the normal cry samples, as we are aware that the number of normal infant cry samples is higher than the pathological infant cry samples, so the results are biased towards normal infant cry class. Hence, the mean classification accuracy is also defined here, which is the average of classification accuracies of the individual class.

Table 6.25. Performance of the features for classification of normal, asthma and HIE infant's cries

S. No.	Feature	Classification Accuracy (in %)
1	Bispectrum	64.95
2.	MFCC	63.88
3.	LPC	67.97
4.	PLPCC	68.15

Confusion matrices are shown in Table 6.26- Table 6.29. From Table 6.26 which represents the classification performance of the PLPCC feature for the three class classification task, we can observe that the PLPCC feature can classify the normal infant's cries correctly with a classification rate of *77.9* %. However, the classification correctness for the pathological cries is very poor. Only *35.7* % of asthma infant's cries and *47.6* % HIE infant cries are diagnosed correctly. Though the misclassification among pathologies is very small, however, misclassified sampled are treated as normal infant sample in this case, which is not required.

Table 6.26. Confusion matrix for the classification of asthma, HIE and normal infant's cries using PLPCC features. Mean classification accuracy is *53.77* %

		Identified as			
		Asthma	HIE	Normal	% Correct classification
True	Asthma	20	1	35	35.7
	HIE	7	49	47	47.6
	Normal	54	35	314	77.9

Table 6.27. Confusion matrix for the classification of asthma, HIE and normal infant's cries using LPC features. Mean Classification accuracy is *62.63* %

		Identified as			
		Asthma	HIE	Normal	% Correct classification
True	Asthma	30	5	21	53.6
	HIE	7	65	31	63.1
	Normal	65	51	287	71.2

Confusion matrix for the classification of normal and pathological cries using LPC-based features is shown in Table 6.27. The LPC-based features are found very poor in the classification of asthma infant cries (53.6%), though its performance for classification of normal and HIE classes is comparatively good which are *71.2* % for HIE and *63.1* % for HIE.

Table 6.28. Confusion matrix for the classification of asthma, HIE and normal infant's cries using MFCC features. Mean classification accuracy is *61.00* %

		Identified as			
		Asthma	HIE	Normal	% Correct classification
True	Asthma	38	1	17	67.9
	HIE	2	49	52	47.6
	Normal	86	45	272	67.5

Table 6.29. Confusion matrix for the classification of asthma, HIE and normal infant's cries using bispectrum features. Mean classification accuracy is *54.1* %

		Identified as			
		Asthma	HIE	Normal	% Correct classification
True	Asthma	9	3	44	16.1
	HIE	3	80	23	77.7
	Normal	44	107	276	68.5

Confusion matrices for the features MFCC and bispectrum-based features are shown in Table 6.28 and Table 6.29. It can be observed that the classification correctness of normal infant cries is similar in both the features, namely, *67.5* % for MFCC and *68.5* % for bispectrum-based features. MFCC is found to perform well in the classification of asthma infant cries with a classification accuracy of *67.9* % and bispectrum-based features are good in classifying HIE infant's cries correctly with a classification accuracy of *77.7* %. Based on the above shown confusion matrices and mean classification accuracy of the features used, it is found that the LPC-based features are better than other features in classifying the samples in the three classes. However, the performance of this feature set is poor in the classification of asthma infant cries.

Assuming that the classification of normal *vs*. abnormal (including HIE and asthma) infant cries is possible with some features and classifiers, we have attempted a classification of asthma and HIE cries in the next Sections of this Chapter. The features used are based on modulation spectrogram, glottal inverse filtering, auditory spectrogram and modified group delay.

6.9 Modulation Spectrogram-Based Features

In *1997*, Greenberg introduced the modulation spectrogram, that display signal in terms of time and slow-varying modulations in a signal (*i.e.*, low frequency modulations) [**180**]. It captures modulation frequencies between *0-8* Hz. Modulation spectrogram features have already been successfully applied for speech recognition and phoneme classification tasks. This feature has shown promising results in these areas.

In this work, we are using modulation spectrogram features with higher-order singular value decomposition (HOSVD) to find the feature vectors for classification of pathologies. This idea was used in the classification of voice disorder (using Massachusetts Eye and Ear Infirmary (MEEI) database) in [**181**]. Here, after application of HOSVD, we are using simple *logarithm* of the features (*i.e.*, log-energy is used as a feature), and then these features are applied to the SVM classifier. However, in [**181**], authors have used a complex method which uses Mel filterbank *smoothing* of the modulation spectrogram with entropy for feature selection.

6.9.1 Modulation Spectrogram

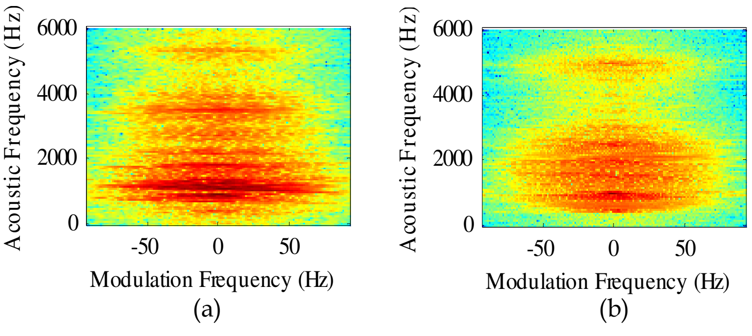
Modulation spectrogram represents the distribution of energy between *modulation* frequency and *acoustic* frequency. The spectrogram is calculated using Short-Time Fourier Transform (STFT) of a signal which gives the distribution of signal between time and acoustic frequency parameters [**180**] and [**182**]. To find the modulation spectrogram, first, STFT of the signal is taken, and then another STFT of magnitude of spectrogram of a speech segment is calculated, *i.e.*,

$$X_m(k) = \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_{I_1}^{kn}, \tag{6.29}$$

$$X_l(k,i) = \sum_{m=-\infty}^{\infty} g(lL-m)\,|\,X_m(k)\,|\,W_{I_2}^{im}, \tag{6.30}$$

where *k* = *0*, 1, 2, …, (I_1-1) and *i* =*0*, 1, 2, …, (I_2-1). *k* is an *acoustic* frequency which is conventional Fourier decomposition of the signal and *i* is *modulation* frequency. *h(n)* and *g(m)* are the analysis windows for *acoustic* frequency and *modulation* frequency, respectively, with hop sizes *M* and *L, respectively*. I_1 and I_2 are the bin sizes in acoustic and modulation frequencies, respectively. $W_{I_1} = e^{-j(2\pi/I_1)}$, $W_{I_2} = e^{-j(2\pi/I_2)}$ and *m* is samples in time-domain. The modulation spectrogram then displays modulation spectral energy $|\,X_l(k,i)\,|^2 \in \mathbb{R}^{I_1 \times I_2}$ in joint acoustic and modulation frequency planes.

In Figure 6.12, modulation spectrograms are shown for both the pathologies under consideration. Figure 6.12 (a) shows the modulation spectrogram of the HIE cry sample. This shows the distribution of energy in both *modulation* and *acoustic* frequencies. The energy is concentrated across *all* acoustic frequency range while its distribution across modulation frequency is in the range of *-90* Hz to + *90* Hz. However, the concentration of energy in modulation spectrogram of asthma cryunit is in the acoustic range which is less than *4* kHz and modulation frequencies of + *50* Hz to *-50* Hz. It can also be observed that amount of energy in asthma cry is lower than the HIE cryunit. Due to these differences in the modulation spectrogram, it can be inferred that it can be a used for classification of asthma and HIE cry samples.



(a) (b)

Figure 6.12. Modulation spectrogram: (a) HIE (b) asthma. In all subfigures, X-axis denotes modulating frequency in Hz, and Y-axis denotes acoustic frequency in Hz.

6.9.2 Experimental Setup

Database: *Corpus II* is used in these experiments. The number of infant cries of asthma samples are *7* cries and of infants suffering from HIE are *16*. One cry per infant was recorded. The cry samples were divided *manually* into cryunits. By dividing the cry samples in cryunits, total *183* cryunits were extracted from asthma cry samples and *216* cryunits were extracted from the HIE cry samples. From these cryunits, modulation spectrogram is extracted and then HOSVD is applied to reduce the dimension of the feature vector. On the reduced feature set, classification performance is evaluated. In this Section, results of three different experiments are reported to show the robustness of the proposed features in the classification of pathological infant cries. One of the experiments shows classification performance when setup is *infant-dependent* (experiment 1) and other shows results when setup is *infant-independent I* (experiment 2). The third experiment shows the robustness of the proposed features in the presence of additive white Gaussian noise (AWGN).

6.9.3 Experiment 1

First of all, voiced regions are selected from the cryunits. For voicing detection, standard energy (i.e., ℓ^2 norm)-based algorithm is used. Each of the cryunit is segmented into frames of *200 ms* duration (modulation spectrogram is *suprasegmental* feature [**181**]). For each cryunit samples, first *10* frames are considered, if the number of frames is more than *10* for a cry sample. For each frame, modulation spectrogram is calculated after subtracting the mean from the samples using *modulation toolbox* [**183**]. We get for each frame a modulation spectrogram of the size *41×150.* These modulation spectrograms are stacked together one after another, to form a *tensor*. On this tensor, we apply HOSVD theorem for its dimension reduction. The size of modulation spectrogram feature for each frame reduces to *10×10.* This reduced feature is then arranged in a form of vector forming a feature vector of size *1×100.* For classification, the *logarithm* of these features is taken. These features are then applied to SVM classifier with radial basis (RBF) function kernel, to find the performance. In order to alleviate data-dependency as classification performance, *4*-fold cross-validation is done on the dataset. In all, we have *1317* feature vectors corresponding to asthma and *1501* feature vectors corresponding to HIE cry samples.

6.9.4 Experiment 2

For all the cryunit signals, voiced regions are identified using the same energy-based algorithm (i.e., ℓ^2 norm). Here, for each cryunit, frames of *200 ms* duration are extracted. If the number of frames per cryunit is more than *25,* only *25* frames are considered. Modulation spectrogram is calculated for each frame. Here, a tensor ($41 \times 150 \times N$) is formed per cryunit ($N \leq 25$) and for features of a cryunit, HOSVD is applied (in order to result features independent of infants). The dimension of the modulation spectrogram feature is then reduced to 10×10. Similar to *experiment 1*, feature vector of 1×100 is formed. A logarithm is taken on the reduced feature set. On the reduced feature set, *4*-fold cross-validation experiment is conducted. Classification accuracy (in %) is obtained using same classifier setup. We have *4550* feature vectors of asthma cry samples and *5357* feature vectors of HIE cry samples.

6.9.5 Experiment 3

Following the experimental setup of experiment *2*, the robustness of the proposed features is shown by adding additive white Gaussian noise (AWGN) to the cry samples before any pre-processing. The signal-to-noise ratio (SNR) considered in this experiment are *5* dB, *10* dB and *15* dB.

6.9.6 Experimental Results

On the specified dataset, experiments were conducted. The experimental results report the average classification accuracy which is defined as the average of the ratio of correctly classified samples to a total number of samples in that class in all validation experiments. In our experiments, we have also tried to find the optimal feature size for classification of pathologies. Results are reported with varying feature dimension. Feature selection is made on the basis of *t*-test scores. Experimental results of classification accuracy with different feature sizes are reported in Table 6.30. It can be seen that as the feature size is increasing classification accuracy also increases. After feature size of *70*, the change in classification accuracy is *not* significant. Here, in our experiment, we have considered *100* as an optimum feature size (and for those confusion matrices are shown in Table 6.31 - Table 6.32). Best classification accuracy is achieved for a feature size of *100* is *87.93 %* in experiment *1* and *76.2 %* in experiment *2*. Experiment *1* shows higher accuracies because in this case, common features of all the infant cries are taken out using HOSVD. However, in practical situations, at a time, cry of a single infant has to be tested, so HOSVD should be applied on individual infant rather than on all infant cries. Moreover, results of experiment *1* are also subjected to number of infants.

Table 6.30. Classification accuracy (in %) of asthma and HIE pathological infant cries with modulation spectrogram features applied on SVM classifier with RBF kernel with parameter $\gamma = 0.01$. Adapted from [**173**]

Feature size	10	20	30	40	50	60	70	80	90	100
Experiment 1	**82.36**	85.34	85.87	86.19	87.04	87.17	87.42	87.64	87.55	**87.93**
Experiment 2	70.79	72.16	72.85	73.41	73.91	74.17	74.65	75.15	75.29	**76.20**

Table 6.31. Confusion matrix for classification of asthma and HIE infant cries in *experiment 1* for feature dimension of *100. Adapted from* [**173**]

		Identified as	
True		Asthma	HIE
	Asthma	90.20 %	9.88 %
	HIE	13.64 %	86.36 %

Table 6.32. Confusion matrix for classification of asthma and HIE infant cries in *experiment 2* for feature dimension of *100*

		Identified as	
True		Asthma	HIE
	Asthma	83.5 %	16.5 %
	HIE	29.93 %	70.07 %

Table 6.33. Comparison of modulation spectrogram features (for feature dimension of *100*) with MFCC for classification of asthma and HIE infant cries

	Modulation Spectrogram	MFCC
Experiment 1	**87.93** %	73.54 %
Experiment 2	**76.23** %	64.43 %

Table 6.34. Comparison of modulation spectrogram (for feature dimension of *100*) features with MFCC for classification of asthma and HIE infant cries with varying SNR

SNR (in dB)	Modulation Spectrogram	MFCC
5	63.98 %	54.19 %
10	72.04 %	54.10 %
15	75.26 %	53.58 %
Clean Signal	**76.23** %	64.43 %

From Table 6.31 and Table 6.32, it can be observed that classification of asthma samples in HIE is smaller than its counterpart. The reason for better classification of asthma samples can be the respiratory problem which causes the energy of cry signal to lie in smaller range (also observed in their modulation spectrograms as shown in Figure 6.12).

A comparison of the proposed method with state-of-the-art feature set, Mel Frequency Cepstral coefficients (MFCC) is reported in Table 6.33. It can be observed that in both the experiments, the performance of modulation spectrogram features is better than the MFCC feature set. Table 6.34 shows the comparison of proposed feature set with MFCC feature set in the presence of additive white Gaussian noise. The observations suggest that in the presence of noise, the performance of MFCC feature set decreases drastically from *64.43* % (Table 6.33) to *54.19* % at *5* dB SNR. However, the change in classification accuracy with change in SNR is *not* significant. Modulation spectrogram is showing comparable results at *15* dB SNR when compared to clean cry samples even at *10* dB SNR, the classification accuracy changes by *-2* % only. The performance of the proposed feature set decreases at very low SNR such as *5* dB, however, the performance is much better than MFCC, state-of-the-art method.

The poor performance of MFCC is an implication from its design. MFCC is designed to mimic human auditory system. In particular, if we cannot identify the sounds auditorily, MFCC also cannot find out the differences in sounds. While the modulation spectrogram-based features captures the low frequency modulations of the sound, which are the source of information in speech. These low frequency variations carry the dynamics of speech production. During speech production, articulators move at the rate of *0-20* Hz. In the case ofinfants suffering from HIE, there is poor neural control over these articulators [**184**]. This results in deviation of articulatory movement which is expected to be captured as well by modulation spectrogram. Moreover, generally the infant data from the hospitals are collected in a noisy environment; it is a better approach to use modulation spectrogram-based features for cry classification because it is a robust feature against additive noise (*i.e.*, under signal degradation conditions).

6.9.7 Summary of Results

In this work, authors have proposed the use of modulation spectrogram features for classification of pathological cries. These features seem to be *robust* in their performance compared to MFCC feature set. The proposed feature uses HOSVD theorem for dimension reduction. It has been observed that with increasing feature size, the classification accuracy improves. It may be possible that feature size of more than *100* may perform even better. However, since the change in accuracy is not very significant, we can consider proposed *1*x*100* as optimum feature size. Since, in hospitals, it is difficult to get a silence environment, the features proposed in medical applications should be robust against noisy conditions. Robustness of the proposed feature set is also shown by adding AWGN. Due to generic nature of the problem, this work can be extended to infants with any mother tongue.

6.10 Glottal Inverse Filtering (GIF)-Based Features

Features derived from the glottal inverse filtering of the speech signal are used for classification of pathological infant cries. Glottal inverse filtering is used to estimate the glottal volume velocity waveform (*i.e.*, the excitation source of voicing for infant cry). Here, GIF is used to separate the glottal excitation source and vocal tract filter. The source and the filter features are used for pathological cries classification. Through the experimental results, the importance of both of these features in cry classification task is investigated. The state-of-the-art feature set, namely, Mel Frequency Cepstral Coefficients (MFCC) is also used to compare the performance of the proposed feature set. GIF has been used widely in speech synthesis application [**185**]. Its application has been demonstrated in the medical field also for voice pathologies identification and classification [**186**].

Glottal inverse filtering (GIF) is a process through which the glottal volume velocity waveform is extracted from the speech pressure waveform. The speech or cry production system is based on the source-filter model. The source of the speech production is the oscillations of the vocal folds which are caused by the airflow provided by the lungs. The rate of oscillations of the vocal folds is determined by the subglottal pressure and mass of vocal folds. Since infants have very small subglottal pressure and less mass of vocal folds, compared to the adults, the distance between two consecutive glottal closure instants (GCIs) is relatively less. Thus, the rate of vibration of vocal folds, *i.e.*, F_0 is higher in infants than the adults. These vocal fold excitations are called *glottal excitations* and are filtered by a *physiological* filter made up of the vocal tract and lips radiation effect (which is typically a highpass filter due to lip radiation). The filtering effects of vocal folds are affected by the articulatory movements of the speakers such as position of tongue and mouth opening. According to these articulators positioning, the formants (*i.e.*, resonance frequencies of the vocal tract) vary and the glottal excitation get affected in its spectrum when filtered by the vocal tract and lip radiation effect [**187**].

GIF method tries to find out the models of the vocal tract and lip *radiation* filter. The inverse models are then used for filtering the speech signal to estimate the glottal excitation waveform. From the *source-filter* theory of speech production, the speech signal can be modeled as the *convolution* of glottal excitation source with the vocal tract filter response and lip radiation filter, *i.e*,

$$S(z) = G(z)H(z)\,L(z), \tag{6.31}$$

where *S(z), H(z), L(z)* and *G(z)* are the *Z*- transforms of the speech signal, impulse response of the vocal tract transfer function, lip radiation effect and glottal volume velocity waveform, respectively. From eq. (6.31), *G(z)* can be computed by speech signal as

$$G(z) = \frac{S(z)}{\mathrm{H}(z)L(z)}. \tag{6.32}$$

6.10.1 Glottal Waveform Estimation

In this work, recently proposed *Iterative Adaptive Inverse Filtering* (IAIF) method is used for the glottal flow waveform estimation [**185**]. This method uses autocorrelation method of linear prediction (LP) modeling of the vocal tract. This has the advantage of the use of an *all-pole* structure which is always *stable*. Apart from this, IAIF is computationally simple and uses only a speech signal as an input. The block diagram of the IAIF method is shown in Figure 6.13 [**185**].

The signal *s(n)* is highpass filtered with a cutoff frequency of *70* Hz to remove low frequency noise signals captured by the microphone. LPC analysis of order *1* gives the combined effect of glottal flow and lip radiation. Through inverse filtering their effect is cancelled at block *3*. At the output of block *4, a* primary estimate of the vocal tract filter is obtained. The final estimate of the impulse response of vocal tract filter is obtained at the output of block *10*. The glottal flow waveform is found at the output of block *12*.

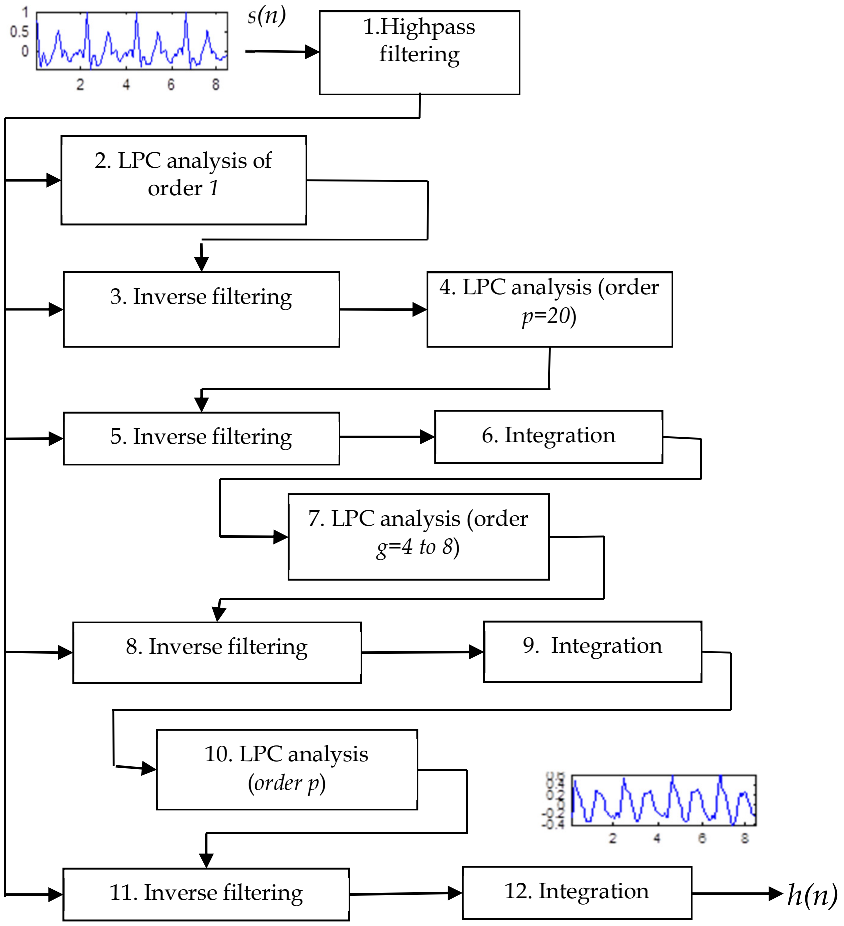


Figure 6.13. Iterative adaptive inverse filtering (IAIF) method for glottal flow waveform estimation. After **[185]**.

6.10.2 Feature Extraction

The voiced speech segments are extracted from the infant cry signal using an energy-based algorithm (*i.e.,* l^2 norm). On this voiced cry signal, Hamming window is applied. The frame duration was kept at *25 ms*. On each frame, IAIF method is applied. This method decomposes the signal into glottal source waveform and all-pole model of the vocal tract, $G(z)$ and $H(z)$, respectively. $H\ (z)$ is made available at the output of block 10, with $p=30$. These coefficients $[H(z)]$ are used as a feature vector. The other feature set considered is the linear prediction coefficients (LPC) of the glottal wave $g(n)$. LP order considered here is *10*. Thus, we have a feature vector [$H(z)$ $G(z)$] (*i.e.*, concatenation of $H(z)$ and $G(z)$). $H(z)$ is of dimension *1* x *30* and $G(z)$ has its dimension *1* x *10*. For each cryunit, first *15* frames were considered and corresponding features were extracted using IAIF.

6.10.3 Experimental Setup and Results

Database: *Corpus II* is used in these experiments. Details of infant cry recordings and cryunits are same as those used in earlier experiments reported in this chapter. For each cryunit, as features were extracted as explained above. On the proposed features, the *classification accuracy* of classification of asthma cryunits and HIE cryunits were found. In our work, classification accuracy is defined as the percentage of correctly classified samples from the total population of one class. Experimental results report the classification performance with $H(z)$ alone as a feature vector, $G(z)$ alone as a feature vector, joint effect of $H(z)$ and $G(z)$ (feature-level fusion of features), MFCC alone for classification, and feature-level fusion of proposed features with MFCC.

Experiment 1: For each of the cryunit, features are extracted for consecutive non-overlapping *15* frames. The feature vector for a cryunit is defined as features of all *15* frames arranged in their order in a single row. For all the cryunits, $H(z)$, $G(z)$ and MFCC features are extracted and saved. In this experiment, *4*-fold cross-validation experiment was conducted on the available feature vectors. SVM classifier is used with the *radial basis function* (RBF) kernel (with $\gamma=0.001$). SVM is a *supervised learning* method which is used for classification and regression. The classification accuracy is reported in Table 6.35.

It can be observed from Table 6.35 that the vocal tract model parameters, *i.e.*, $H(z)$ perform better for asthma and HIE cry classification.

MFCC performs better than the proposed feature set. However, the fusion of MFCC with proposed features outperforms to MFCC. Though the joint performance of *G(z)* and *H(z)* features is better than MFCC, their fusion with MFCC is performing better than the individual MFCC and features derived from GIF in classification. Fusion of MFCC with *H(z)* and *G(z)* features, increases the classification accuracy to *91.42 %* compared to *86.9 %* classification accuracy of MFCC alone. Thus, this finding indicates that proposed features, namely, *G*(*z*) and *H*(*z*) capture complementary information from infant cry signal than the MFCC alone.

Table 6.35. Classification accuracy (in %) with different features when features are arranged in the form of a vector for a cryunit for the classification of asthma and HIE infant cries

Features	Feature size	Classification accuracy (in %)
H(z)	*1* x *450*	85.12
G(z)	*1* x 15*0*	78.35
H(z)+ *G(z)*	*1* x *600*	87.14
MFCC	*1* x 180	86.9
MFCC+*G(z)*	*1* x 330	89.17
MFCC+ *H(z)*	*1* x 210	90.93*
MFCC+ *H(z)*+*G(z)*	*1* x 780	89.90
MFCC+ *H(z)*+*G(z)*	*1 x 500*	91.42*

*after feature selection using *t*-test

Experiment 2: The experiment was conducted using the hold-out method. Out of *6* asthma cries, *4* infant cries were used for training and *2* for testing. Similarly, in HIE, *14* infant cries were used for training and *2* infant cries were used for testing. In training, *128* cryunits of asthma and *169* cryunits of HIE cry samples were taken and remaining cryunits were used for testing. As described for experiment *1*, features were extracted and classification performance of the features was tested on SVM classifier with RBF kernel. Experimental results are reported in Table 6.36 and Table 6.37.

It can be observed from Table 6.36 and Table 6.37 that classification performance of *H(z)* feature vector is better than glottal source parameters *G(z)* and MFCC. Table 6.37 shows the results of *experiment 2*, when *t-test* is applied for feature selection. Maximum classification accuracy achieved in HIE and asthma cry classification is *77.31 %*.

Table 6.36. Classification accuracy (in %) with different features when mean of the features is taken for a cryunit using holdout method for the classification of asthma *vs.* HIE infant cries.

Features	Feature size	Classification accuracy (in %)
H(z)	*1* x *450*	74.22
G(z)	*1* x 150	59.79
H(z)+ *G(z)*	*1* x 600	69.07
MFCC	*1* x 180	71.13
MFCC+*G(z)*	*1* x 330	73.19
MFCC+ *H(z)*	*1* x 630	74.22
MFCC+ *H(z)*+*G(z)*	*1* x 780	75.25

Table 6.37. Classification accuracy (in %) for HIE *vs.* asthma infant cries with different features when mean of the features is taken for a cry using holdout method with feature dimension reduction. Adapted from [**188**]

Features	Feature size	Classification accuracy (in %)
H(z)	*1* x *200*	76.28
G(z)	*1* x *150*	59.79
H(z)+ *G(z)*	*1* x *150*	77.31
MFCC	*1* x *180*	71.13
MFCC+*G(z)*	*1* x *100*	78.35
MFCC+ *H(z)*	*1* x *50*	77.31
MFCC+ *H(z)*+*G(z)*	*1* x *150*	78.35

Experimental results show that in both the pathologies, the vocal tract functioning gets affected, which is reflected in *H(z)*. In addition, vocal folds remain unaffected in these two pathologies or get affected in a similar manner that it cannot work as a feature vector in infant cry classification. To find out this effect, the glottal flow waveforms of both the pathologies should be compared with the normal infant cry. Fusion of proposed features with MFCC performs better than the individual features indicating that proposed features *H*(z) and G(z) captures complementary information compared to MFCC alone. In particular, information which is captured by MFCC is *supplemented* by the excitation *source* or *system* features, and their combined effect is performing better than any of the proposed feature set. *H(z)* feature represents the vocal tract filter. In the case ofasthma disease, there is blockage of airways, however, HIE does not affect the respiratory system. These effects are reflected in the features *H(z)*. While in both the pathologies, the suffering infant has difficulty in breathing. In asthma, it is due to blockage of the vocal tract and in HIE, it is due to oxygen deficiency to the brain. Hence, in both the cases, glottal excitation waveform will be distorted in comparison to normal infant's glottal flow waveform (GFW). The cry waveforms and their corresponding extracted glottal flow waveforms for asthma, HIE and normal infant cry samples are shown in Figure 6.14 and Figure 6.16.

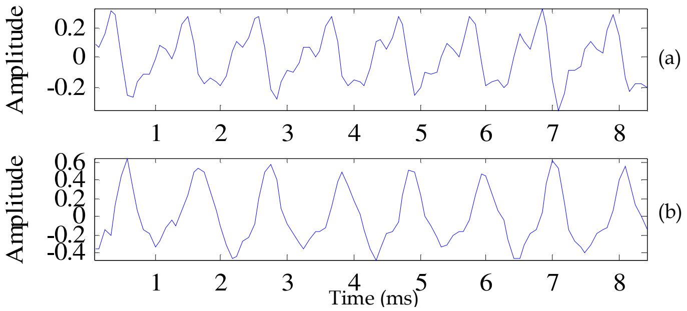


Figure 6.14. (a) Time-domain (b) glottal flow waveform of an *asthma* cry segment.

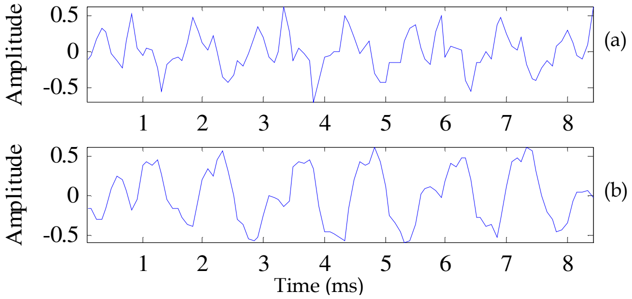


Figure 6.15. (a) Time-domain (b) glottal flow waveform of a *HIE* cry segment.

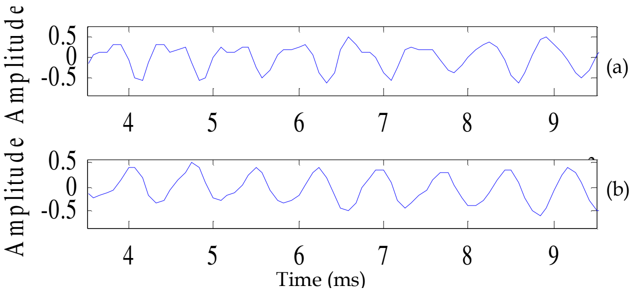


Figure 6.16. (a) Time-domain (b) glottal flow waveform of a normal cry segment.

6.10.4 Summary of Results

From the experimental results, it can be observed that the feature *H(z)* is more powerful in the classification of HIE and asthma pathologies. However, the glottal excitation source information *G(z)* performs poorly in the classification of these two pathologies. Fusion with MFCC feature set improves the classification performance. Classification performance can be further increased by defining classification accuracy as number of infants diagnosed correctly to the total number of infants. The diagnosis for an infant can be judged on the basis of maximum number of cryunits assigned to one of the pathological class. With this classification criterion, classification accuracy can increase further. In our case, because of smaller number of participants (*i.e.*, infants), we have not applied this criterion. However, for the said experiment, we are getting *100 %* classification accuracy.

6.11 Auditory Spectrogram-Based Features

It is well known that the human auditory system (HAS) can perceive the sounds and make distinction between different voices due to pitch, loudness and timbre which are basic elements of the speech signal. Pitch is related to the frequency of the signal while loudness signifies the amplitude of the speech sound (recently, it is found that loudness is associated with the strength of epoch [**150**]). Timbre is the spectral variations with time which we perceive to make differences in different sounds [**189**]. The cochlear model resembles the human auditory model and it works as follows:

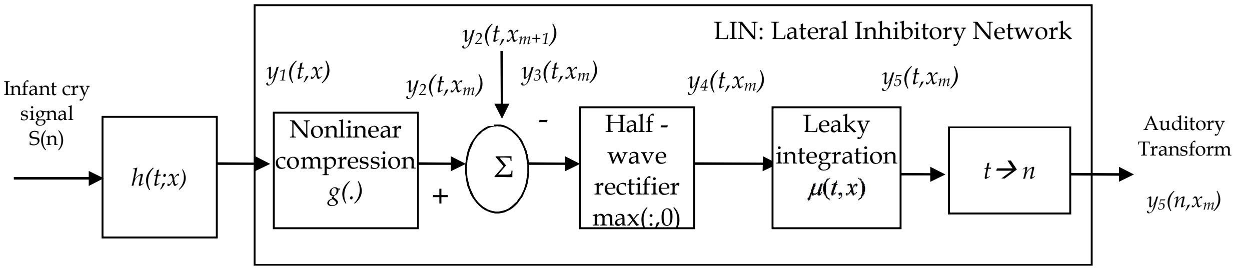


Figure 6.17. Block diagram of the auditory (cochlear) model. After [**189**].

The sound wave impinges upon the outer ear and they are directed towards the inner ear. These travelling waves cause vibrations in the fluid which is filled in the cochlea. The vibrations cause displacement in the *basilar* *membrane* (BM) which is inside the cochlea. The basilar membrane has *20,000* hair-like nerve cells. These nerve cells differ in length and also have different degrees of resiliency to the fluid that passes over them. As the motion wave passes through the fluid, these hair cells are set into the motion. Each hair cell has a natural sensitivity to a particular frequency. When this frequency matches the frequency of vibration of the acoustic travelling wave, then this nerve produces an electrical signal. This electrical signal is carried by the auditory nerve fibers to the auditory cortex of brain where we perceive the sound. Temporal fluctuations in any given fiber are limited to frequencies below *4-5* kHz due to the lowpass filtering effect of the hair cell membranes. Above these frequencies, the auditory nerve indicates the presence of a particular frequency by a steady increase in firing rate [**189**].

The sound signal which is *1-D* pressure *vs*. time waveform is converted to *2-D* acoustic spectrogram when the sound signal goes through different transformation by middle and inner ear. The auditory spectrogram represents the distribution of energy with respect to time and frequency where the frequency is logarithmically distributed. The cochlear model is shown in Figure 6.17 and consists of three stages, *namely,* analysis, transduction and reduction as discussed in [**190**].

Analysis Stage

In the analysis stage, the response of the basilar membrane (BM) is modeled for the sound signal *s(t).* Eq. (6.33) represents the analysis stage, *i.e.*,

$$y_1(t,x) = s(t) *_t h(t,x), \tag{6.33}$$

where *h(t,x)* denotes the impulse response of the filter located at *x* (*x* is spatial location away from the cochlea) and $*_t$ is the *convolution* in time-domain. The analysis stage is implemented by *128* overlapping subband filters with constant *Q* (i.e., quality factor) bandpass filters. The center frequencies of the bandpass filters are logarithmically distributed. The position and frequency are related by $x = \log_2\left(\frac{f}{f_0}\right)$ in octaves relative to f_0 which is some reference frequency (*e.g., 1* kHz). The logarithmically transformed frequency is called *tonotopic* frequency.

Transduction Stage

The transduction stage is modeled in eq. (6.34). The equation calculates the response of the hair cell which incorporates fluid cilia coupling, compressive ionic channels and membrane leakage, *i.e.,*

$$y_2(t,k) = g(\partial_t y_1(t,x)) *_t w(t). \tag{6.34}$$

The temporal derivative describes fluid cilia coupling which is computationally equivalent to the pre-emphasis on the incoming signal. The function *g(.)* (*i.e.,* nonlinear compression) model the nonlinear channel through the hair cell. *w(t)* is a lowpass filter which filters out all responses beyond *4* kHz and is used to model the leakage of the cell membrane. In speech signal, most of the energy lies below *4* kHz, its role is minor and hence, can be ignored at this stage.

Reduction Stage

In the reduction stage, the lateral inhibitory network (LIN) was divided into three steps, tonotopic derivative (eq. (6.35)), half wave rectification (eq.(6.36)) and leaky integration (eq. (6.37)), *i.e.,*

$$y_3(t,x) = \partial_x(y_2(t,x)) *_t v(t), \tag{6.35}$$

$$y_4(t,x) = \max(y_3(t,x),0), \tag{6.36}$$

$$y_5(t,x) = y_4(t,x) *_t \mu(t,x). \tag{6.37}$$

The derivative ∂_x simulates the lateral interaction among LIN neurons. The spatial filter *v(.)* models the local smoothing due to finite spatial extent of the lateral interactions. The half wave rectifier *max(.,0)* mimics the positive nature of LIN neurons. Finally, a temporal integration window $\mu(t,\tau) = e^{-t/\tau}u(t)$ is applied to model the slow adaptation of central auditory neurons. The time-frequency representation $y_5(t,x)$ is called the *auditory spectrogram*.

6.11.1 Auditory Spectrogram *vs.* Short-Time Fourier Transform Spectrogram (STFT)

Spectrogram of a speech signal gives a visual representation of the variation of its energy with time over the frequency spectrum. Spectrogram of a signal *x(n)* is defined as follows:

$$X(m,k)=\sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_{I_1}^{kn}, \tag{6.38}$$

where $k = 0,1, \ldots, (I_1-1)$, I_1 is the number of bins to be considered where k is an *acoustic* frequency which is conventional Fourier decomposition of the signal. *h(n)* is the analysis windows with hop sizes *M*. $W_{I_1} = e^{-j(2\pi/I_1)}$ and m is samples in time-domain.

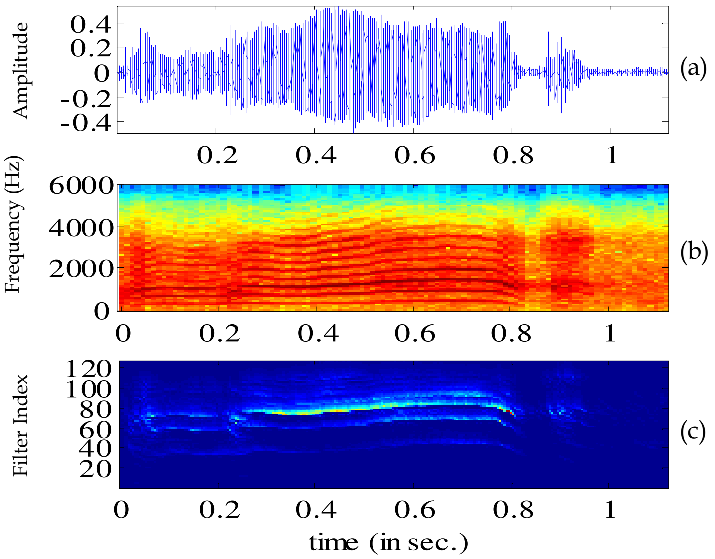


Figure 6.18. (a) Cry segment of a HIE cry, (b) its spectrogram and (c) corresponding auditory spectrogram.

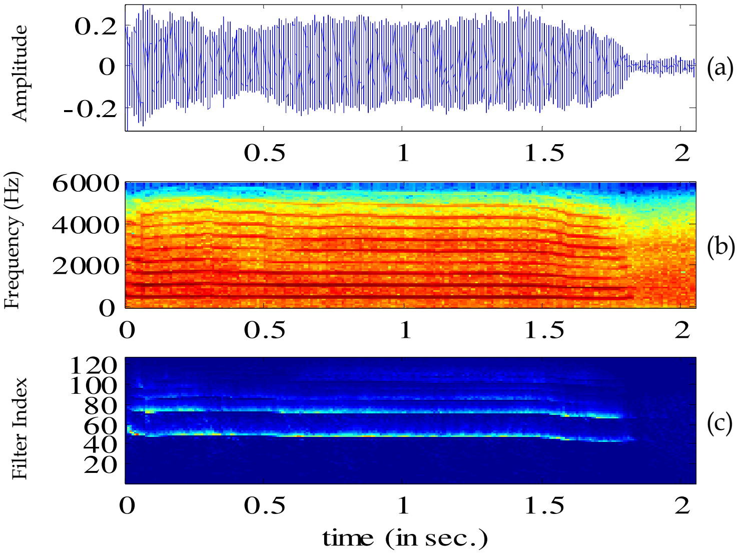


Figure 6.19. (a) Cry segment of an asthma cry, (b) its spectrogram and (c) corresponding auditory spectrogram.

Spectrogram thus displays the energy distribution of the signal in time and frequency-domains. Figure 6.18 and Figure 6.19 show the examples of spectrograms and auditory spectrograms for HIE and asthma cry samples, respectively. In Figure 6.19 (c), Y-axis shows the frequencies distributed logarithmically and X-axis is time (in sec). In spectrogram, it can be observed that energy is distributed across *all* frequency components (*i.e.*, harmonics). However, the auditory spectrogram retains the formant structure and removes the noise energy concentration, thereby improving the visualization of the distribution of energy in limited and useful frequency regions only.

6.11.2 Feature Extraction

Initially, all infant cry samples are divided into cryunits manually as shown in Figure 6.20. For each of the cryunit, the auditory spectrogram is calculated after pre-processing. In the pre-processing stage, the cry signals are passed through a *4th* order lowpass Butterworth filter with a cutoff frequency at *3* kHz. It removes the highpass noise in the signal. Then features are extracted from estimated auditory spectrogram. The auditory spectrogram is a *2-D* feature set calculated over complete cryunit duration.

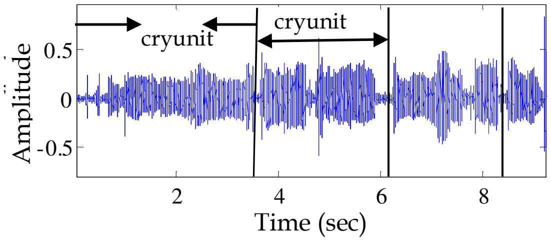


Figure 6.20. Selection of cryunit from infant cry recording.

To extract the useful features from the auditory spectrogram, the features are summed over time and frequency-axis, *i.e.*,

$$X(\omega) = \sum_t X(t,\omega), \tag{6.39}$$

$$X(t) = \sum_\omega X(t,\omega), \tag{6.40}$$

where the auditory spectrogram of a cryunit is represented by $X(t,\omega)$, ω is frequency-axis, t is time-axis. For each cryunit, the auditory spectrogram will have a different number of frames depending on the length of the cryunit. To have similar length of auditory spectrograms, the number of frames has been fixed to *250*, which in turn yields auditory spectrogram of *128* x *250* for each of the cryunit. If a cryunit has smaller number of frames, then zeros are padded in the trailing side of feature vectors. After feature reduction using the proposed method, we get feature vectors of size *1* x *128* for each of the cryunit. After getting these feature vectors, the logarithm is taken which serve as a feature vector for the classifier. These features are then applied to a support vector machine (SVM) classifier for training and testing. Here, in our experiment, radial basis function (RBF) kernel is used for SVM classifier.

6.11.3 Experimental Setup and Results

Database: In our experiments, *Corpus II* is used. The number of infant cries of asthma samples are *7* cries and of infants suffering from HIE are *14*. By dividing the cry samples in cryunits, total *250* cryunits were extracted from asthma cry samples and *295* cryunits were extracted from the HIE cry samples.

While conducting the experiment, it has been assumed that classification of asthma and HIE infant cries from normal infant cries is available. In order to further classify the pathologies, here, an attempt is made towards the classification of asthma and HIE infant cries. For each of the cryunit extracted from the two classes, the feature vectors are estimated using the proposed method. To find the significance of the proposed feature set, classification accuracy is estimated. In these experiments, classification accuracy is defined as the percentage of correctly classified samples out of total population. To quote the *statistical significance* of the experimental results, performance is compared with the MFCC feature set with *10*-fold cross-validation is performed. In *10*-fold cross-validation experiments, infants in HIE and asthma classes are randomly divided in train and test datasets. *10* different combinations of infants are considered in training and test datasets. For each combination of train and test dataset, classification accuracy is found using SVM classifier with RBF kernel function.

Table 6.38. Distribution of cryunits in training and test datasets for *10*-fold cross-validation experiment

	Asthma		HIE	
Group	Training	Testing	Training	Testing
1	185	63	163	130
2	159	89	155	137
3	107	141	150	142
4	193	55	168	124
5	199	49	178	114
6	125	123	143	149
7	137	111	178	114
8	125	123	144	148
9	137	111	199	94
10	163	85	141	151

In asthma, out of *7* infants, *4* infants are kept in training and *3* in testing and in HIE class, out of *14* infants, *8* infants are kept in training and *6* are kept in testing. The statistics of cryunits in all *10* combinations are shown in Table 6.38. The number of cryunits of each infant is not the same. Cryunits in a cry which are longer than *0.75* seconds are considered as useful cryunit in this experiment because in longer cryunits effect of harmonics and formants effect are clearly visible.

Table 6.39. Classification accuracy (in %) in *10*-fold cross-validation experiment. Adapted from **[191]**

Group No.	$X(t)$	$X(\omega)$	$\log(X(t))$	$\log(X(\omega))$	MFCC	Combined feature vector
1	32.64	32.64	60.61	90.15	83.93	90.67
2	39.38	39.38	67.35	89.823	93.08	92.47
3	50.17	50.17	50.18	85.86	74.91	80.212
4	30.73	30.73	69.27	83.24	84.91	87.71
5	30.06	30.06	69.9	81.59	80.36	82.22
6	54.77	54.77	54.77	87.87	88.97	89.71
7	50.67	50.67	50.66	93.33	95.11	95.55
8	54.62	54.62	54.61	90.04	89.66	89.67
9	45.85	45.85	45.85	85.85	81.46	87.32
10	36.02	36.02	63.98	88.98	88.98	89.83
Average	42.29	42.29	58.7	87.67	86.13	88.54

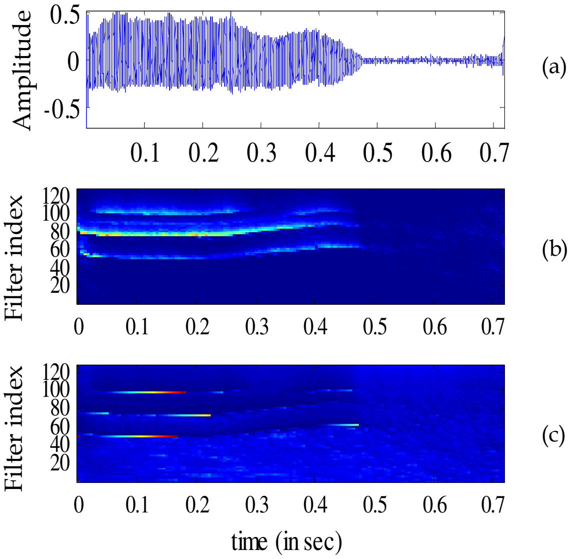
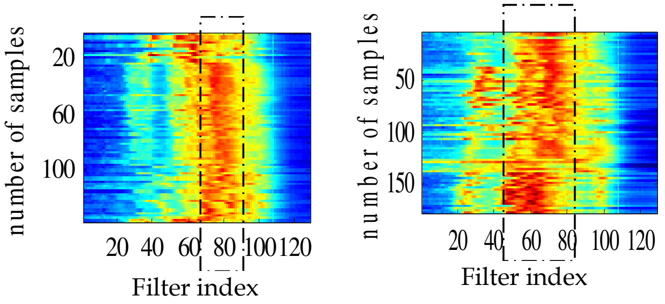


Figure 6.21. Difference in proposed feature and auditory spectrogram (a) cry segment, (b) auditory spectrogram and (c) auditory spectrogram after logarithm.

Classification performance of the proposed feature set is indicated in Table 6.39, the performance of the MFCC features and fusion of proposed features with MFCC is also shown. It can be observed that same classification results are obtained with $X(t)$ and $X(\omega)$ features. When these features are used, mostly samples were classified either as HIE or as asthma cry samples that resulted in the values shown columns *2* and *3* of the Table 6.39 (as the dataset is imbalanced dataset). Results shown in Table 6.39 show that summation of auditory spectrogram along time dimension is *better* than its summation along frequency-axis, for the proposed classification task. Summation along frequency-axis gives temporal variations in amplitude. Hence, does not give a good performance. On the other hand, summation along time-axis captures the formants and dominant frequency (harmonic) information, which can be a good parameter of difference in the two pathologies, *namely*, asthma and HIE. To find out the reason of the better performance of the proposed feature set, refer Figure 6.21. In Figure 6.21(b) and Figure 6.21 (c), differences can be observed when logarithm is applied on the auditory spectrogram. After application of logarithm on the features of eq. (6.39)-(6.40) and taking its absolute value, formants and dominants harmonic frequencies become clearer and these can contribute to the higher performance of proposed features.



(a) (b)

Figure 6.22. Distribution of energy in different frequency components of all cry samples for (a) asthma and (b) HIE. The dotted rectangle indicates the region where energy is higher.

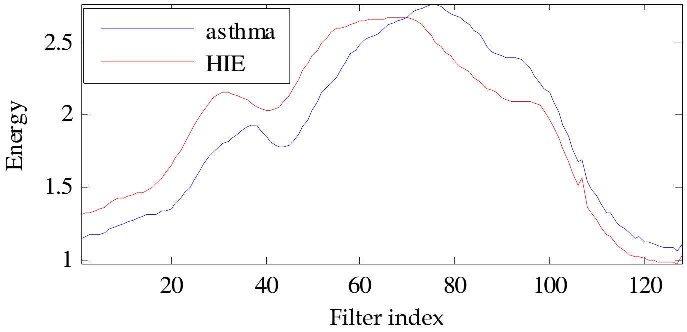


Figure 6.23. Energy (mean) distribution across frequencies of asthma and HIE cry samples.

The differences in the proposed features between two pathologies can be found in Figure 6.22. It can be seen in Figure 6.22 (a) that in asthma the distribution of energy lies between filter inddex *60-80* (shown by dotted lines) and their corresponding frequencies. However, in HIE, energy is distributed between filter index *40-80* (shown by dotted lines) and their corresponding frequencies, in low frequency range (below frames *40*) energy is also distributed. The difference in energy distribution over different frequencies is also shown in Figure 6.23. It can be observed that in lower frequencies, energy is higher in HIE cry samples whereas in the case ofasthma, energy is higher in higher harmonic frequencies.

6.11.4 Summary of Results

In this work, features derived from auditory spectrogram are proposed and these features show better classification performance than the state-of-the-art method. The proposed feature set works closely on the principle of how our brain perceives the speech/cry sound. When we listen to any sound, a *2-D* image is created in our brain. According to the pattern generated, we perceive the information carried by the sounds waves. This *2-D* pattern or auditory spectrogram represents variations in amplitude with time as well as the frequency.

In spectrogram, which is also time *vs*. frequency representation of speech, we do not get better intelligibility because in that we use Fourier transform of the signal to convert temporal information in spectro-temporal information. Fourier transform does not consider the sensitivity of human ear in different frequency bands. Hence, all frequencies get equal importance. In MFCC, this frequency sensitivity of ear to different frequency components is taken care by Mel scale filtering. However, the nonlinearity of the hair cell channel is not considered in MFCC. This can be the reason for the better performance of the proposed feature set than MFCC. If we can find a feature which can classify HIE and asthma infant cries from normal infant cries, then we can have a promising application for asthma *vs.* HIE classification.

6.12 Group Delay-Based Features

Recently, modified group delay features have shown promising results for automatic speech recognition (ASR) task [**192**]. In this work, features are derived from the modified group delay function. These features are then used to classify asthma and hypoxy ischemic encephalopathy (HIE) infant cries.

6.12.1 Modified Group Delay Function

Information carried in the Fourier transform phase of the signal is conveyed implicitly by the group delay function. To make this representation more effective and useful, instead of group delay, modified group delay function is used which captures the vocal tract system-related information of the vocal system. Modified group delay function has been proved effective in the semi-automatic *segmentation* of speech and features derived from the modified group delay function have been applied for language identification (LID), automatic speech recognition (ASR) and speaker identification (SID) [**192**], [**193**], [**194**]. Normally, the information in the speech signal is represented by the short-time Fourier transform (STFT). In that method, the magnitude of the FT is considered for feature extraction and phase of FT is ignored. The significance of phase for several speech applications has been shown in recent studies [**195**]. It has been proved that in presence of significant phase distortions, the recognition of speech by the human ear is very poor. The information in the FT phase can be extracted by the negative derivative of the unwrapped Fourier transform phase, *i.e.*, group delay function, $\tau(\omega)$ is given by,

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega}, \tag{6.41}$$

where $\theta(\omega)$ is the *unwrapped* Fourier transform phase function. The negative sign in the expression for $\tau(\omega)$ is used to realize the output of a causal system, *i.e.*, output at the filter (having group delay characteristics as $\tau(\omega)$) will be there only when the first input is applied and not before the input is applied. The group delay function can be computed from the discrete-time FT (DTFT) as given below [9]:

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{6.42}$$

where $X(\omega)$ and $Y(\omega)$ are the FT of the signal $x(n)$ and $nx(n)$, respectively. The subscripts R and I represents the real and imaginary parts of the DTFT, respectively. The group delay function requires that the signal be minimum phase signal (*i.e.*, all the poles and zeros of the system lie inside the unit circle). The group delay function becomes spiky on or near the unit circle in z-domain [**192**]. To remove the excitation source information, the denominator term $|X(\omega)|$ can be replaced by its cepstrally smoothed envelope. The cepstrally smoothed envelope is given by $S_c(\omega)$. The modified group delay is represented as [**193**]

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S_c^{\,2}(\omega)}. \tag{6.43}$$

To reduce the dynamic range of modified group delay spectrum and peaks at the formant locations, two parameters α and γ were introduced in [**193**]. The new modified group delay is defined as

$$\tau_c(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau(\omega)|)^{\alpha}, \tag{6.44}$$

where $\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S_c(\omega)^{2\gamma}}$ and the parameters range from $0 < \alpha \leq 1$, $0 < \gamma \leq 1$.

6.12.2 Feature Extraction

Features are extracted from the proposed method as shown in the Figure 6.24. Initially, the cry recording from an infant is divided into cryunits. From each cryunit, mean is subtracted. The cry signal is filtered with a *4th* order

Butterworth lowpass filter to remove high frequency noise. Since most of the signal information is contained in signal below *4* kHz, the filter cutoff frequency is taken as *4* kHz. The cry signal is segmented into non-overlapping frames of *10 ms* duration. For each of the frame, modified group delay as mentioned in eq. (6.44) is computed. In our experiments, we have kept $\gamma = 1$ and $\alpha = 0.1$. For the same frame, MFCCs are also calculated. One feature vector is extracted for a cryunit. Mean is taken over all the frames of a cryunit (same is done for MFCC feature vector as well). After getting a modified group delay feature vector for a cryunit, its discrete cosine transform (DCT, type II, *i.e.*, DCT-II) is taken. The modified group delay cepstra is taken as a feature vector for the proposed work. DCT is applied on group delay coefficients to decorrelate the feature vectors. In this work, we have taken *30*-point DCT. Experiments were conducted to find the number of suitable points for DCT. Based on these experiments, *30* DCT coefficients are considered optimum for the classification task, which is considered as a feature vector for infant pathology classification work.

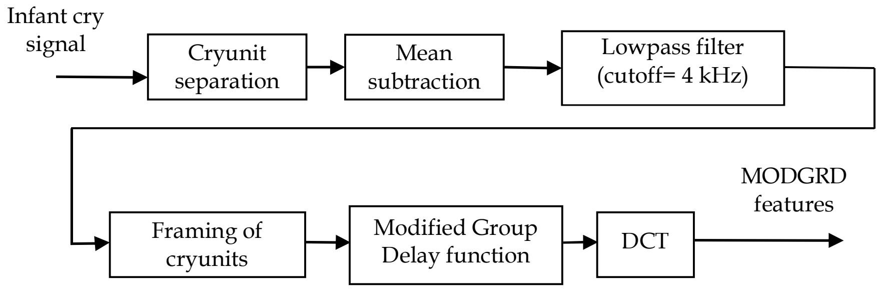


Figure 6.24. Block diagram for extraction of features from modified group delay function.

6.12.3 Database Used

The database used in this experiment is same as the one used in Section 6.11, for the classification task using auditory spectrogram.

6.12.4 Experimental Results

For each of the cryunit, features are extracted as explained above. These features are then applied to an SVM classifier for training and testing. In our experiments, we have considered polynomial kernel and RBF kernel for SVM classifier. The results of classification are reported in Table 6.40- Table 6.41. Classification accuracy (in %) is defined as the ratio of a total number of samples correctly classified to the total number of samples multiplied by *100*.

Table 6.40. Classification accuracy (in %) with MFCC and proposed MODGRD features with different degrees of polynomial kernel function applied to SVM classifier.

Kernel order	MFCC	MODGRD
3	76.99	78.76
4	72.56	84.96

Table 6.41**.** Classification accuracy (in %) with MFCC and proposed MODGRD features with radial basis function (RBF) kernel applied to SVM classifier. Adapted from [**196**]

Value of γ (kernel parameter)	MFCC	MODGRD
0.1	78.76	76.11
0.5	79.64	88.49
0.25	79.64	90.25
0.125	78.76	81.41

Table 6.40 shows the classification accuracy (in %) for asthma and HIE cry samples classification with both MODGRD and MFCC feature sets using SVM classifier with polynomial kernel function. It can be observed that the highest classification accuracy of *84.96* % is achieved with a polynomial kernel of degree *4* with the MODGRD features. Even with the polynomial order of *3* kernel, gives better performance with proposed features compared to MFCC. Table 6.41 shows the classification performance of the proposed task with RBF kernel function. It can be observed that the class separability is highest with kernel parameter γ =*0.25*. Highest classification accuracy achieved for the modified group delay-based features is *90.25* %. Confusion matrices for the classification of asthma and HIE infant cries for MFCC and proposed features with different kernels are shown in Table 6.42 - Table 6.45.

Table 6.42**.** Confusion matrix with proposed feature set using *4th* order polynomial kernel function

		Identified as	
		Asthma	HIE
True	Asthma	39	16
	HIE	1	57

Table 6.43**.** Confusion matrix with MFCC feature set using 4^{th} order polynomial kernel function. Adapted from [**196**]

		Identified as	
True		Asthma	HIE
	Asthma	30	25
	HIE	6	52

Table 6.44. Confusion matrix with proposed feature set using γ = *0.25* with RBF kernel function

		Identified as	
True		Asthma	HIE
	Asthma	46	9
	HIE	2	56

Table 6.45**.** Confusion matrix with MFCC feature set using γ = *0.25* with RBF kernel function

		Identified as	
True		Asthma	HIE
	Asthma	37	18
	HIE	5	53

Table 6.46. Classification of cryunits of individual infants in the two classes using proposed and MFCC features

Features			MODGRD		MFCC	
Kernel function			Polynomial kernel (p=4)	RBF	Polynomial kernel (p=4)	RBF
Number of cryunits		Total	Correctly identified cryunits			
Asthma	Infant 1	22	9	16	11	14
	Infant 2	22	15	21	13	16
	Infant 3	11	7	9	6	7
HIE	Infant 1	8	8	8	8	8
	Infant 2	10	10	10	8	10
	Infant 3	25	25	23	12	22
	Infant 4	6	6	6	6	5
	Infant 5	9	9	9	8	8

It can be observed that the correct classification of HIE infant cries is better than asthma infant cries in both the features along with any classifier kernel function. However, the modified group delay-based features are able to capture differences in asthma and HIE infant cries and performing better than MFCC features. Details of classification for an individual infant's cryunits are reported in Table 6.46. It can be seen that using the proposed features with SVM classifier employing RBF kernel has better distinction of pathologies compared to the MFCC. In HIE classification, both the features are performing equally well. However, in asthma classification, proposed features classify more number of cryunits to asthma class compared to MFCC. It assures the correct identification of pathology. If *70* % of the cryunits are considered as a threshold for identification of pathologies, then, in MFCC, all the three infants are not sure to have asthma disease, whereas proposed features give assured results for *2* infants to have asthma out of *3* infants. However, with RBF kernel function, all infants are correctly identified for their pathologies with the same threshold of *70* %. Figure 6.25 shows the cluster plots of features derived from modified group delay function and MFCC features, respectively.

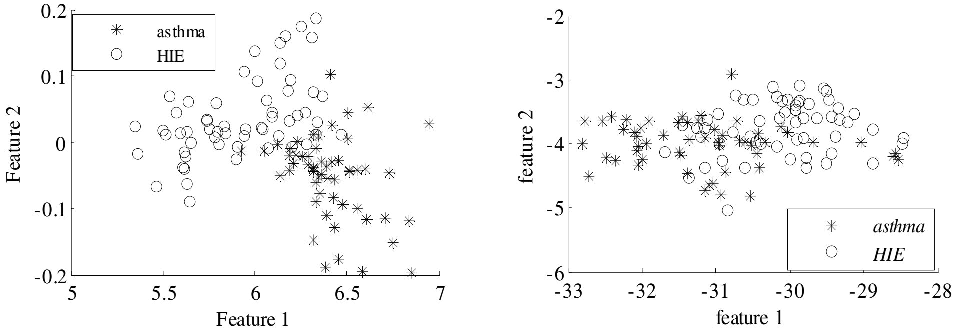


Figure 6.25. Clustering of feature vectors for the two classes in two-dimensional (*i.e.*, *2-D*) feature space (a) MODGRD and (b) MFCC.

To compare the performances of the proposed features and MFCC, their class separability distances are calculated using Bhattacharya bound. The Bhattacharya distance is a measure of similarity of two discrete or continuous probability distributions. It is used to measure separability of classes in classification. Bhattachraya distance between two classes is given by **[197]**

$$D(p,q)=\frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2}+\frac{\sigma_q^2}{\sigma_p^2}+2\right)\right)+\frac{1}{4}\left(\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_q^2}\right), \tag{6.45}$$

where *p* and *q* are the two classes, σ_p and σ_q are the variances of *p* and *q* classes, respectively. μ_p and μ_q are the means of classes *p* and *q*, respectively.

For better visibility of the curves, only first *12* coefficients of the feature vector are considered in the plot. Figure 6.26 shows that the class separation of the proposed features is higher than the MFCC feature set.

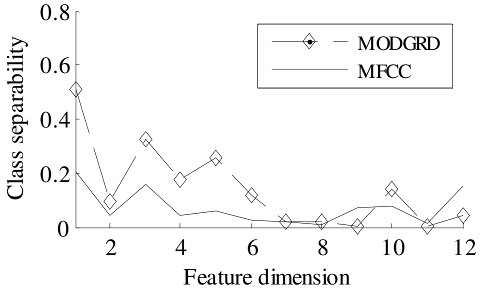


Figure 6.26**.** Class separability of MFCC and proposed features. After [**196**].

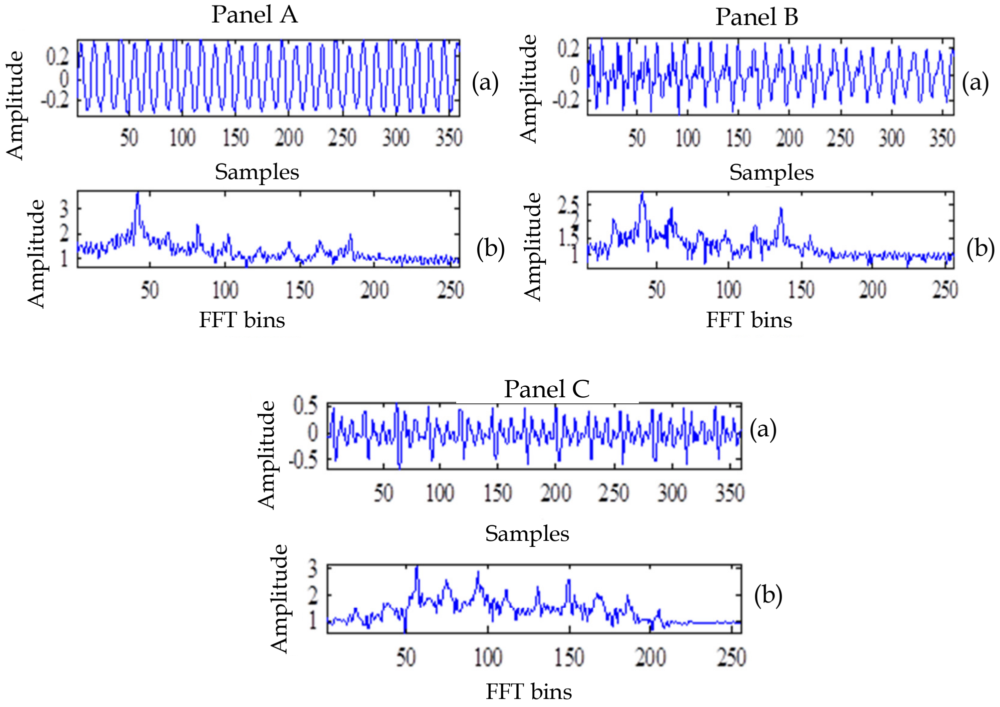


Figure 6.27. Modified group delay function for normal (Panel A), HIE (Panel B) and asthma (Panel C) infant cries. In all subfigures, (a) corresponds to infant cry signal and (b) corresponds to its group delay function.

Figure 6.27 shows the signal waveform and their respective modified group delay functions for the three cases, normal, HIE and asthma infant cries. It can be observed that in the case of normal and HIE infant cries, we get a higher number of peaks of group delay function. However, in asthma infant cries, we get well separated peaks in the group delay function. These peaks in the group delay function correspond closely to formants of the infants. Hence, there may be changes in the formants and their harmonics of the infants corresponding to a particular pathology.

6.12.5 Summary of Results

In this Section, features derived from the group delay are proposed for the classification of asthma and HIE infant cries. It has been observed that the modified group delay-based cepstral features outperform the MFCC feature set which has been the state-of-the-art method in infant cry classification. Both the features are capable of capturing the information in HIE cry samples. However, classification of asthma pathology is better in the proposed features. The improved performance of the proposed features may be due to its property that it captures the Fourier transform phase information of the signal, while MFCC is derived from the magnitude spectrum of the FT, ignoring the phase spectrum completely. This difference in two features suggests that in asthma patients, phase variations are very frequent. These fast variations in phase of the signal may be introduced by the *blocked airways*. It has also been observed that the selection of kernel function is also important in the classification work. For the proposed features, best results are obtained with the *4*th degree polynomial function whereas for MFCC, best results are obtained with *3rd* degree polynomial function kernel. However, the proposed features are giving good performance at *3rd* degree kernel function as well. In this chapter, the best performance of both the features is compared irrespective of the kernel functions. In this work, we have assumed that classification of normal and pathological cry samples is possible with some features and after this classification, we can use the proposed method for classification of asthma and HIE diseases in infants.

6.13 Classification of Normal *vs*. Deaf Infant's Cries

In this Section, classification of normal and deaf infant cries is attempted. The classification is performed on the *Corpus III* (Baby Chillanto Database). In the classification task, the cry segments or cryunits which are *1* sec. long are used. The statistics of these cryunits are given in Table 6.47. The dataset is divided in train and test datasets as follows:

Table 6.47. Distribution of cryunits in normal and deaf class

Class	No. of cryunits
Normal train	446
Normal test	61
Deaf train	335
Deaf test	544

6.13.1 Experimental Setup and Results

For each of the cryunit, the cry signal is divided into smaller duration frames and for each frame LPC, MFCC and bispectrum features are extracted. These are used to classify normal and deaf infant cries. For the estimation of bispectrum, the indirect method is used and the feature extraction methods as defined in Section 6.4 and Section 6.4.4 are used. The performance measure used here is the classification accuracy (in %), which is defined as the ratio of the sum of true positive and true negative to the total number of samples multiplied by *100*. The classification performance is evaluated after *4*-fold cross-validation. In all the experiments, SVM classifier is used with RBF kernel. The parameter selected in RBF kernel is $\gamma = 0.01$ and LIBSVM tool is used. The experimental results are shown in Table 6.48 - Table 6.49.

From Table 6.48, it can be observed that the highest classification accuracy is obtained with MFCC feature set, *i.e.*, *84.98* %. When the amplitude is normalized to ±*1* and the sampling frequency is changed to *16* kHz, the classification accuracy improves. For bispectrum-based features, it improves from *76.9* % to *88.98* % and for MFCC feature-set the improvement is from *84.98* % to *92.38* % as shown in Table 6.48 and Table 6.49. Effect of variation of frame length is observed on the classification accuracy. It is observed that in the case of MFCC feature-set, the classification accuracy increases from *92.38* % to *93.5* % with a sampling frequency of *16* kHz, however, it decreases in the case of bispectrum features when the frame size is increased from *30 ms* to *100 ms* as shown in Table 6.49 and Table 6.51. The same trend is observed with *8* kHz sampling frequency as well (Table 6.50 and Table 6.51).

Table 6.48. Classification accuracy (in %) with different features for the frame size of *30* ms and sampling frequency of *8* kHz.

S. No.	Feature Set	Feature Dimension	Classification Accuracy (in %)
1.	MFCC	1 x 39	**84.98**
2.	LPC	1 x 11	**83.00**
3.	Diagonal slice of bispectrum estimated using lag *8*.	1 x 128	76.90
4.	Diagonal slice of bispectrum estimated using lag *4*.	1 x 128	76.90
5.	Arithmetic, geometric and harmonic means of bispectrum	1 x 192	76.29

Table 6.49. Classification accuracy (in %) with different features for the frame size of *30* ms and sampling frequency of *16* kHz.

S. No.	Feature Set	Feature Dimension	Classification Accuracy (in %)
1.	MFCC	1 x 39	**92.38**
2.	Bispectrum cumulants	1 x 130	83.58
3.	Diagonal slice and peaks locations of bispectrum estimated using lag *8*.	1 x 130	**88.98**
4.	Bispectrum SVD	1 x 128	81.02

Table 6.50. Classification accuracy (in %) with different features for the frame size of *100* ms and sampling frequency of *8* kHz.

S. No.	Feature Set	Feature Dimension	Classification Accuracy (in %)
1.	MFCC	1 x 39	**93.06**
2.	Diagonal slice of bispectrum estimated using lag *8*.	1 x 128	76.16
3.	Diagonal slice of bispectrum estimated using lag *4*.	1 x 128	76.9
4.	Diagonal slice of bispectrum estimated using lag *12*.	1 x 192	76.24
5.	Bispectrum SVD lag *8*	1 x 128	72.45
6.	Bispectrum SVD lag *12*	1 x 128	71.77

*SVD= singular values of the bispectrum taken as a feature vector.

Table 6.51. Classification accuracy (in %) with different features for the frame size of *100* ms and sampling frequency of *16* kHz

S. No.	Feature	Feature Dimension	Classification Accuracy (in %)
1.	MFCC	1 x 39	**93.5**
2.	Diagonal slice of bispectrum estimated using lag 8.	1 x 128	**83.6**

From the experimental results, it is observed that MFCC feature set performs well for the classification of normal and deaf infant cries. The best classification accuracy is obtained with the sampling frequency of *16* kHz for the cry frame duration of *100 ms*, which is *93.5 %*. For this experimental setup, the mean MCC, SP and SN scores are *0.84*, *0.86* and *0.975*, respectively. The reason can be this feature set is modeled for the perception of the speech. As the distinction between deaf and normal infant cries can be made auditorily, this feature set is able to classify the normal and deaf infant cries.

6.14 Chapter Summary

In this chapter, classification of normal *vs.* pathological infant cries is attempted using bispectrum-based features. The performance of the direct and indirect methods of bispectrum feature extraction methods is compared. In particular, indirect method of bispectrum estimation is found to perform slightly better than direct method. A comparison of proposed feature extraction technique HOSVD with other existing feature extraction methods and robustness of proposed bispectrum features under noisy conditions is also shown and it is found to perform better than the state-of-the-art methods. Apart from this, classification of asthma and HIE infant cries is shown. For the classification task, four different features based on modulation spectrogram, glottal inverse filtering, auditory spectrogram and modified group delay are used. After comparing the results of these features, it has been found that the performance of modified group delay features is relatively better than other feature extraction methods. The reason is the ability of the MODGRD features to capture the nonlinearity introduced in the Fourier transform phase spectrum of the signal. The difference in nonlinearity in the two pathologies is illustrated and has been found that in asthma it is due to blocked airways. In the classification of normal and deaf infant cries, MFCC feature vector is found to perform better than the other features. In the next chapter, the work presented in this thesis is concluded along with a discussion on future research directions.

Chapter 7.

Conclusions and Future Work

The aim of the thesis was to investigate the scope of signal processing methods for the analysis and classification of infant cries. Infants are equally important part of our community and thus, fulfillment of their needs and protection needs to be taken care by the elders. Because, infants cannot communicate their requirements using a language, a translator is required to enable the caretaker for understanding the requirements of the infants. This is a socially-relevant research problem and the objectives set for the thesis were as follows:

1. To create a database for analysis and classification of infant cries,
2. To identify the signal processing challenges associated with infant cry signal processing,
3. To analyze the different cry types and identifying the acoustic descriptors of the various cry types,
4. Identifying the infant's pathological health from the cry signal,
5. Classifying the pathologies from the infant cry signal,
6. Study of high risk infants who are at high risk of SIDS.

7.1 Conclusions

After performing experiments in the direction to achieve the set objectives mentioned above, following conclusions are made:

1. Author of this thesis has collected *Corpus I* from Civil Hospital, Ahmedabad, in a real-life setting. This also resulted in analysis of various ideal characteristics of the corpus, metadata preparation, ethical issues, consent form for parents, *etc.*

2. It was shown with the experiments that the acoustics of infant cry signal changes with several factors. Important among these factors are age, weight and reason of crying of an infant. The age factor becomes more important in the case of prematurely born infants and in such cases, GA of the infant becomes a dominating factor in the infant cry analysis. In the first three months of the birth (*i.e.*, neonatal period), the development of the CNS and coordination among motor activities (*i.e.*, motor control mechanism) is very fast making and hence, this group of infants is difficult to study. In other words, we can say that the infant cry analysis can be divided into three different groups that are (a) birth cry analysis (b) neonatal or newborn cry analysis, (c) infant cry analysis (*1-12* months).
3. ANOVA analysis of basic acoustic descriptors of the infant cry such as minimum F_0, maximum F_0, mean F_0 and durational features suggests that the F_0 is higher in pathological infant cries. Pain cries also have higher F_0 compared to the hunger cries. The higher pitch (F_0) values are the parameters associated with the urgency of the call. Duration of the cry is not useful parameter pathology identification. However, it is associated with the reason of crying. In hunger cries, the duration is longer than the pain cries.
4. Higher unvoicing ratio in infant cries may not be attributed to the presence of pathology. This decision also depends on in the age of the infant. The spectrographic studies show that the *dysphonation* which is captured in the unvoicing ratio is present in the pathological infant cries. However, results reported in our analysis shows that in normal neonates, the unvoicing ratio is higher than the pathological cries. In infants, dysphonation shows a lack of integration and coordination in the vocal production system while in adults it is associated with the irregularities and poor CNS control of the vocal folds.

5. In neonates, reduction in the unvoicing ratio is the indicator of the maturity of CNS. In neonates, the energy of the infant cry signal lies in *2-4* kHz band of frequencies. In the case of pain cries, the concentration of the energy is found in higher frequency bands.
6. For classification of asthma and HIE infant cries, modulation spectrogram-based features are used and found to perform well in clean as well as in the noisy environment. Modulation spectrogram-based features capture the low frequency variation of the signal which is due to articulators' movement in the case of infants. This results in good classification accuracy of these features.
7. Auditory spectrogram considers the sensitivity of human ear for different frequencies. It is observed from this feature set that the pattern of distribution of energy in two pathologies, are different. In the case of asthma, energy lies in high frequency bands, while in HIE, the energy of auditory spectrogram lies in low frequency ranges making the distinction or discrimination better in the two pathologies.
8. Applying the glottal inverse filtering (GIF)-based approach for classification of asthma and HIE pathologies in infants from their infant cry signals, it is observed that the features representing the vocal folds vibration cannot classify the cry signals of asthma and HIE. It indicates that in both the pathologies, either the vocal folds are not at all affected or both pathologies affect the vocal tract in a similar manner. However, vocal tract parameters are good in classifying the two pathologies. In asthma, there is swelling in vocal tract which narrows the vocal tract and results in asthma and hence, it results in changes in the vocal tact parameters. In HIE, the brain gets damaged due to oxygen deficiency which probably results in an abnormality in the medulla oblongata which is known to control the breathing phenomenon and hence, poor control is found over the vocal tract which results in a problem in breathing. In both the cases, glottal flow

waveform will be distorted as compared to normal infant cry signal glottal flow waveform because there is problem in breathing to infants.

9. Group delay features show that because of the presence of the pathology, FT phase characteristics of the infant cry signal also changes. These changes in the FT phase of the signal shows the irregularities in the infant cry sound produced which in turn reflect the poor mechanism of vocal folds in the pathological infants.
10. Analysis of infants at high risk shows that these infants emit cries that are louder than the normal infant cries. However, in infants suffering from respiratory distress energy-level in a cryunit is lower than the normal infants. It is because the infants with respiratory distress (RDS) have shorter cries and poor respiratory control.

7.2 Scope of Future Research Directions

The research area of infant cry analysis using signal processing method is new and a very little work is done in this direction by several researchers mainly due to several signal processing challenges/difficulties associated with the infant cry signal. It has a huge scope of work in signal processing and infant cry analysis and classification.

7.2.1 Signal Processing Methods

a. It is shown in Chapter 4 that the choosing correct order of LP is a difficult task in infant cry analysis. As there is a slight increase in the LP order, the LP spectrum matches with the pitch (F_0) contour of the system which is not desired. Since the LP order is found to be different for different cry signals because of a large range of fundamental frequencies, work can be done in order to develop an algorithm to estimate the correct LP order which can work for infants, children and adults such that the same system design can be used for many purposes.

b. A similar problem is found in the cepstrum analysis of the infant cry. Though using the cepstrum analysis on infant cry signal, we can get a good estimate of F_0 and formants, however, the problem lies when this method is used for the estimation of the harmonic-to-noise ratio (HNR). In the recent trends in HNR measurements, cepstrum is widely used because it gives good accuracy. In this method, the *lifter* is used to separate the excitation source and system information. The accuracy of the results depends on the accuracy of the separation of these two components. Since lifter size is very small in the case of infants and changing the lifter size by a single sample (for *12* kHz sampling frequency), makes a large change in the HNR values. This requires the research in the direction of estimation of appropriate lifter size.

c. Fundamental frequency (F_0) measurement is a well-known challenging problem in infant cry analysis. All the methods which use a cutoff values of F_0 for estimation of peaks in the signal derived from the original cry signal, cannot work well in this area. Because the frequency range is quiet large in infant cry signals which ranges from *200* Hz to *1* kHz (sometimes more than this as well). Hence, deciding a threshold for peak-picking is a difficult task.

d. Formant estimation in the case of hyperphonic sounds is a difficult task which needs to be addressed. In infants, the first formant (F_1) lies around *1100* Hz and hyperphonic sounds also have fundamental frequency of vibrations more than *1* kHz. Thus, the harmonics of the F_0 makes it difficult to distinguish formants from the harmonics.

e. Estimation of glottal flow waveform (GFW) is a challenging task for infants. Many methods have been tried during this thesis work such as zero frequency filtering (ZFF)-based method, glottal inverse filtering (GIF) method, Hilbert transform-based method. However, none of the methods is giving correct glottal waveform which makes estimation of glottal parameters such as open quotient (OQ), closed quotient (CQ) difficult.

7.2.2 Infant Cry Analysis Methods

a. In the present work, few features are used for the analysis of the infant cry signals. More features such as *prosodic* features can be used to analyze the infant cry signal which will give a better understanding of the underlying mechanism of articulators in the cry production mechanism.
b. Analysis can be done for various infant cry types to relate the signal processing methods to understand the physiology of the cry production and central nervous system (CNS) which can help the pediatricians to understand the anatomy of the infants.
c. Study can be carried out in the direction of understanding of the maturity of the physiological system in the infants using the infant cry analysis. This requires long-term follow up of the many infants which is a difficult task.
d. If for the pathological cases, long-term follow up is possible, then research can be done in diagnosing the reason of the pathology using signal processing methods. It requires coordination of various professionals.
e. Cry research in mature infants (*8* months - *2* years) can help in understanding the language acquisition in infants, effect of multilingual environment on infant's vocal learning, identifying the reason of late learning in late learners.
f. In infants with speech disorders, or hearing loss, identifying the symptoms and pathology in early days can help in getting late in learning of such infants. Work can be done in this direction.

g. Work can be done in the direction of identifying the infants from their cry. Such a system if developed can help a working mother in taking care of their infants in daycare as well.

h. Amongst all these, a standard database can be created for the infant cry sounds. While developing the database, care must be taken to consider all possible factors which affect the cry performance of an infant. This can help in developing the interest of the researchers and comparing the results of several algorithms used for the same purpose.

Appendix A

Participant Information Form

1.	Name of Infant			
2.	Name of Mother			
3.	Name of Father			
4	Infant's date of birth		Premature/ full term	
5.	If Premature, Gestation Age (GA)			
6.	Type of delivery	Normal/ C-section		
7.	Weight at present		Weight at birth	
8.	Mother's Details	Housewife/ Working (at the time of pregnancy)		
		Use of tobacco/ alcohol		
		Weight/height		
		Qualification		
		Age		
9.	Any disease to father			
	Any disease to mother			
10.	Details of siblings			
11.	Income Group	Less than 2 Lac / more than 2 Lac		
12.	Food habits	Vegetarian/ Non Vegetarian		
13.	SIDS History			
14.	Infant health condition	Normal/ pathological		
15.	Type of Cry	Pain/ Hunger/ Pleasure/ Birth		

16.	Diagnosis by doctor	

Signature and Name of Doctor with Hospital:

Date: Folder Number: File Number:

Destination file path (where data is stored for future use):

Appendix B

Consent Form

Date:

I, the undersigned____ consent voluntarily to enroll my ward in the research project conducted by Ms. Anshu Chittora who is a doctoral student at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar. This study will be carried out for the study of infant cry for pathology identification and classification at DA-IICT, Gandhinagar. I am informed about the procedure of this study. I know that it is a non-commercial and educational study. I am informed that the identity of myself and the results are kept confidential throughout the study and the results will be used for research purpose only. I am not provided any remuneration for participation in this study. I also know that I am free to withdraw the consent any time during study, and it will not affect the treatment rendered to me. I give consent on my own free will.

Signature __________________

Name and relation with participant:

Address:

Appendix C

Analysis of Variance

Analysis of Variance (ANOVA) is used to test for the significant differences between means. The result produced by ANOVA is same as *t*-test if the groups under study are two and the samples are *independent*. The name ANOVA came from the fact that in order to compare statistical significance of means, variances of the two groups are compared.

The underlying three assumptions in ANOVA are as follows [**198**]:

1. Observations are chosen randomly and independently from the population;
2. Within each group, the samples are normally distributed, and
3. The variance of all the groups is same.

When there is only one variable then the ANOVA is called one-way ANOVA [**199**].

Symbols used in the analysis are:

k = number of groups,

n_i= the sample size taken from the i^{th} group,

x_{ij} = the j^{th} response obtained from the i^{th} group,

$\bar{x}_i$ = the sample mean of the responses from the i^{th} group, *i.e.*, $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$,

s_i = the sample standard deviation from the i^{th} group, *i.e.*, $s_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$,

n = total number of samples, *i.e.*, $n = \sum_{i=1}^{k} n_i$,

$\overline{x}$ = the mean of all responses irrespective of group, *i.e.*, $\overline{x} = \sum_{ij} x_{ij}$.

Thus, the total amount of variability among the observation is given by the sum of squares total (*SST*).

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 \ . \qquad \text{(C. 1)}$$

This total variability has two sources of variability, namely, variability between the groups (SSB) and variability among groups (SSW). These are defined as follows:

$$SSB = \sum_{i=1}^{k} n_i(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})^2 \ \text{and} \ SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 \ , \qquad \text{(C. 2)}$$

Here, *SST*=*SSB*+*SSW*. If the variability between groups is large compared to variability within groups, then the data suggests that the means of the population from which the data are drawn are significantly different. That indicates the *F*-ratio, $Fratio = \frac{MSB}{MSW}$ where *MSB* is the $\frac{SSB}{k-1}$ and *MSW* is $\frac{SSW}{n-k}$. ANOVA table is given in Table C.1:

Table C.1. ANOVA Table. After [**200**]

Source	*SS*	*df*	*MS*	*F*
Between samples	*SSB*	*k-1*	*MSB*	$\frac{MSB}{MSW}$
Within samples	*SSW*	*n-k*	*MSW*	
Total	*SST = SSB+SSW*	*n-1*		

If *F* is large, variability between samples is large compared to variability between groups. Variable '*df* ' defines the degree of freedom. The obtained value of *F*-ratio from the Table C.1 is compared with the value of *F*-ratio obtained from the *F*-Table for the given degrees of freedom. If the obtained *F* value from ANOVA analysis is bigger than the value obtained from the *df*, then the null hypothesis is rejected. It indicates that there is at least one of the variable among groups which contributes significant differences within the groups.

Appendix D

Support Vector Machine (SVM) Classifier

In the classification tasks shown in this thesis, classification accuracy is determined using support vector machine (SVM) classifier with radial basis function (RBF) kernel on the reduced feature set. LIBSVM tool is used in our work [**177**]. Support vector machine (SVM) algorithm was introduced by Boser, Guyon and Vapnik in 1992 [**201**]. SVM is a supervised learning method which is used for classification and regression. SVM constructs a hyperplane or a set of hyperplanes in the high-dimensional space that can be used for the classification task. The optimal hyperplane is the one, which gives maximum distance between the nearest training samples. Nearest training samples are called the *support vectors*. If the datasets are inseparable, then kernel functions are used to nonlinearly map the input data to high-dimensional space. In this high-dimensional space, the dataset becomes linearly separable (due to Cover's theorem) [**202**]. Kernel functions make this transformation simple by utilizing inner product with the input data. The kernel function which is commonly used is radial basis function (RBF) kernel because of their localized and finite responses. The kernel functions used with SVM classifier are shown in Table D.1.

Table D.1. Various kernel fuctions used in SVM

S. No.	Kernel Name	Kernel function $K(X_i, X_j)$
1.	Linear	$X_i.X_j$
2.	Radial Basis Function (RBF)	$\exp(-\gamma(\| X_i - X_j \|^2)$
3.	Polynomial	$(\gamma X_i.X_j + c)^d$
4.	Sigmoid	$\tanh(\gamma X_i.X_j + c)$

where $K(X_i, X_j) = \phi(\mathrm{X}_i).\phi(X_j)$, *i.e.*, the kernel represents the dot product of the input data points mapped into high-dimensional data plane by the transformation ϕ [**203**]. An example of using SVM classifier in our work is shown in Figure D.1. In this figure, classification of infant cries and adult speech using energy-based features is shown (described in Section 5.4). Here, the features E_{1n} and E_{2n} are used to train the SVM classifier. Using these features, selection of support vectors and decision plane is shown in Figure D.1

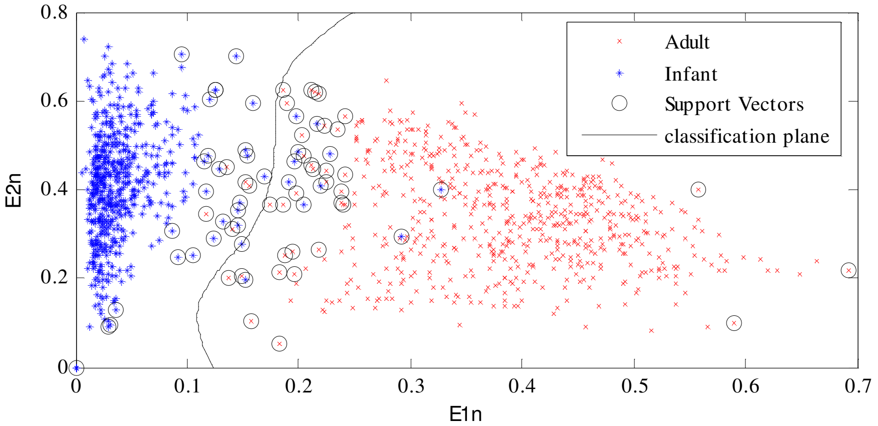


Figure D.1. Separation of data using SVM classifier with RBF kernel.

SVM has the advantages of being efficient in high-dimensional spaces and saving memory space (because it requires a subset of training samples). It is also effective when the number of dimensions is greater than the number of samples. However, it does not perform well when the dataset is large, because it requires large time to train. For inseparable or noisy data, its classification performance is very poor.

Appendix E

Hidden Markov Model (HMM)

HMM has been used widely in speech applications such as automatic speech recognistion (ASR) and word recognition [**204**]. HMM is used to learn the states through which a sequence has undergone from the sequence of emission. Analysis of the HMM seek to recover the sequence of states from the observed data [**205**], [**206**], [**207**]. The outcome of an experiment is called the observation symbol o. The notations used are:

N= number of states,

M= number of observation symbols,

T=length of observation sequence,

i_t= states in which we are at time t,

$V = \{V_1, V_2, V_3, \ldots\}$ discrete set of possible observation symbols,

$\pi = \{\pi_j\}$ probability of being in state j at the beginning of the experiment, *i.e.*, at $t=1$, *i.e.*, $\pi_j = P(i_1 = j)$, $A = \{a_{rj}\}$ The probability of being in state j at time ($t+1$) given that we were in state r at time t. We assume that a_{rj} are independent of time, i.e., $a_{rj} = P(\mathrm{i}_{t+1} = \mathrm{j} \mid \mathrm{i}_t = \mathrm{r})$.

$B=\{b_j(k)\}$, $b_j(k)=P(V_k \text{ at } t \mid i_t = j)$ is the probability of observing the symbol V_k given that we are in state j.

O_t= observation symbol at instant t.

The HMM is denoted as $\lambda = (A, B, \pi)$.

For the model to be of use to practical applications, the three basic problems must be solved. These three problems are:

Problem 1. Given the observation sequence O= { $o_1, o_2, o_3, \ldots, o_T$ } and the model $\lambda=(A,B,\pi)$, how to calculate $P(O/\lambda)$?

Problem 2. Given the observation sequence O= { $o_1, o_2, o_3, \ldots, o_T$ }, and the model $\lambda=(A,B,\pi)$, how to choose a corresponding state sequence $I=i_1,i_2,i_3,\ldots,i_T$ so that $P(O,I|\lambda)$, the joint probability of the observation sequence and the state sequence given the model is maximized?

Problem 3. How do we adjust the HMM model parameters $\lambda=(A,B,\pi)$ so that $P(O|\lambda)$ is maximized?

Problem 1 is the *evaluation* problem (compute the probability of observed sequence produced by the model), problem 2 attempts to identify the correct state sequence and problem 3 attempts to optimize the model parameters to best describe how an observation sequence was generated.

Solution to problem 1: To calculate the probability of the observation sequence O= { $o_1, o_2, o_3, \ldots, o_T$ }, given the model $\lambda=(A,B,\pi)$, *i.e.*, $P(O|\lambda)$. The most straightforward way to find $P(O|\lambda)$ is to find $P(O|I,\lambda)$ for a fixed state sequence $I=i_1,i_2,i_3,\ldots,i_T$, then multiply it by $P(I|\lambda)$ and then sum up over all possible *i's*. We have

$$P(O|I,\lambda)=b_{i_1}(o_1)b_{i_2}(o_2)\ldots b_{iT}(o_T), \tag{E. 1}$$

$$P(I|\lambda)=\pi_{i_1}a_{i_1i_2}a_{i_2i_3}\ldots a_{i_{T-1}i_T}. \tag{E. 2}$$

Hence, we have

$$P(O|\lambda)=\sum_I P(O|I,\lambda)P(I|\lambda). \tag{E. 3}$$

This results in heavy calculations of the order of $2T\,N^T$. To solve the problem 1, another possible methods are forward procedure and backward procedure.

Solution to problem 2: Here, we need to find a state sequence $I = i_1, i_2, i_3, \ldots, i_T$ such that the probability of occurrence of the observation sequence $O = \{ o_1, o_2, o_3, \ldots, o_T \}$ from this state sequence is greater than that from any other state sequence. In other words, the problem is to find I that will maximize $P(O, I \mid \lambda)$. Viterbi algorithm is used to solve this. It is an inductive algorithm in which at each step we keep the best path.

Solution to problem 3: this problem deals with training the HMM such that it encodes the observation sequence in such a way that if an observation sequence having many characters similar to the given one be encountered to it should be able to identify it. Baum-Welch algorithm is used for this.

E.1 HTK-HMM

Hidden Markov Model tool kit (HTK) is a toolkit for building Hidden Markov Models (HMMs). HMM is used to model any time series. There are two processing stages involved in HTK, first, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions, second, unknown utterances are transcribed using the HTK recognition tools.

E.2 The HMM Parameters

An HMM consists of a number of states. Each state j has n associated observation probability distribution $b_j(o_t)$ which determines the probability of generating observation o_t at time t and each pair of states r and j has an associated transition probability a_{rj}. In HTK, the entry state 1 and the exit state N of an N-state HMM are *non-emitting*.

For a 5- state HMM there are three emitting states and have probability distributions associated with them. The transition matrix for this model will have 5 rows and 5 columns. Each row will sum to one except for the last row which is always zero because no transitions are allowed from the output state.

HTK is concerned with the continuous density models in which each observation probability distribution is represented by a mixture Gaussian density.

Before any training or recognition can be done with HTK, we have to setup the required data compatible to HTK. HTK configuration file which in this case contains the following parameters:

SOURCEKIND = WAVEFORM, SOURCEFORMAT = NIST, TARGETKIND = MFCC_Z_E_D_A, LOPASS = 300, HIPASS = 3400, NUMCHANS = 26, NUMCEPS = 12, ENORMALIZE = T, CEPLIFTER = 22, TARGETRATE = 100000, WINDOWSIZE = 250000, ZMEANSOURCE = T, USEHAMMING = T, PREEMCOEF = 0.97

HTK can use separate label files for each speech file, however, more efficient is the usage of so called Master Label Files (MLF) that store the label files independently of the location of the wave files into one common structure. The HTK tool HLEd is a general purpose label editor. It can be used to arrange individual label files into an MLF.

Appendix F

Mel Frequency Cepstral Coefficients (MFCC)

The MFCC feature set is used widely in automatic speech and speaker recognition tasks. MFCC was introduced by Davis and Mermelstein in 1980 [**208**]. It has been a state-of-the-art method in ASR as well in infant cry analysis methods. When we speak, the excitation source signal is filtered by the shape of the vocal tract and these changes in the vocal tract are reflected in the power spectrum of the speech signal. The MFCC features capture these changes in the power spectrum, thereby the vocal tract characteristics of the speech sound. MFCC takes into consideration the human perception of the speech by considering the Mel scale of the frequencies rather than linear scale.

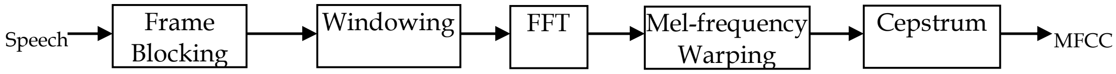


Figure F.1. Block diagram for MFCC feature set extraction.

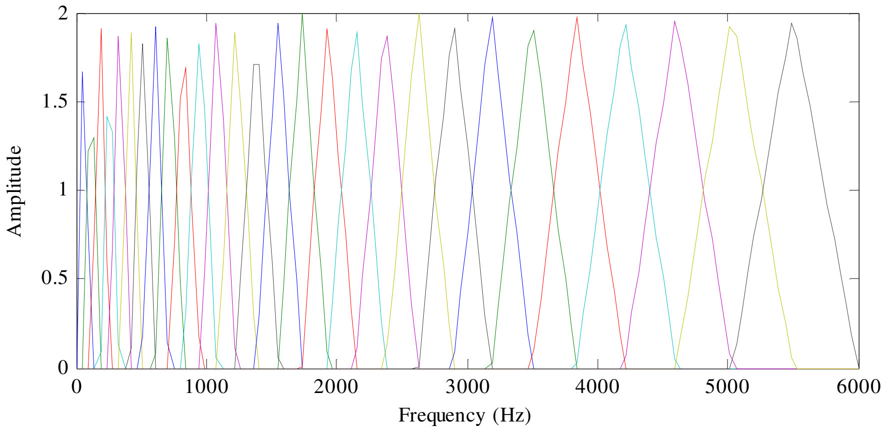


Figure F.2. Mel-filterbank used in infant cry analysis in this thesis.

The steps followed in MFCC estimation shown in Figure F.1 are as follows:

1. Segment the speech/signal into small frames of duration *10-30* ms,

2. For each frame, calculate the power spectrum,

3. Apply Mel filterbank (as shown in Figure F.2) to the power spectra and sum the energy of each of the filter output,

4. Take logarithm of the filterbank energies,

5. Apply DCT to the filterbank energies derived in step *4,*

6. Keep DCT coefficients *2-13* and discard others.

References

[1] B.M. Lester and C.Z. Boukydis, "No language but a cry," in *Nonvedrbal vocal communication*, H. Papousek and U. Jurgens, Eds., 2008, ch. 8, pp. 145-173.

[2] Merriam-webster. merriam-webster.com. [Online Available] http://www.merriam-webster.com/dictionary/pathological {Last accessed on May 15,2016}

[3] O. Wasz-Hockert, K. Michelson, and J. Lind, "Twenty five years of Scandinavian cry research," in *Infant Crying- Theoritical and Research Perspective*, Barry M. Lester and C.F. Zachariah Boukydis, Eds. New York and London: Plenum Publising corporation, 1985, ch. 4, pp. 83-104.

[4] H. Farsaie Alaie and C. Tadj, "Cry-based classification of healthy and sick infants using adapted boosting mixture learning method for Gaussian mixture models," *Modelling and Simulation in Engineering*, vol. 2012 (2012), no. 1, pp. 1-10, 2012.

[5] M. Z. Mohd. Ali, W. Mansor, Y.K. Lee, and A. Zabidi, "Asphyxiated infant cry classification using simulink model," in *8th International Colloquium on Signal Processing and its Applications*, Melaka, 2012, pp. 491-494.

[6] M. Hariharan, L. S. Chee, and S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural networks," *J. of Medical Systems*, vol. 36, no. 3, pp. 1309-1315, September 2010.

[7] Babypod [Online Available]. http://www.dailymail.co.uk/sciencetech/article-3386181/Bizarre-babypod-tampon-speaker-play-music-unborn-children.html {Last accessed on May 15,2016}

[8] Cryingbebe. [Online Available]. https://play.google.com/store/apps/details?id=com.aco.cryingbebe {Last accessed on May 15, 2016}.

[9] C. Gregoire, The Huffington post in association with Times of India Group. [Online]. http://www.huffingtonpost.com/entry/robotic-baby-infants-moms_56041996e4b0fde8b0d18334?section=india {Last accessed on May 15,2016}

[10] R. Gelin et al., "Towards a storytelling humanoid robot," in *AAAI Fall Symposium: Dialogue with Robots*, Virginia, USA, 2010, pp. 137-138.

[11] B. M. Lester, "Inroduction- There's more to crying than meets the ear," in *Infant crying-Theoritical and Research Perspective*, Barry M. Lester and Zachariah C.F. Boukydis, Eds. New York and London: Plenum Press, 1985, pp. 1-27.

[12] E. Gustafsson, F. Levrero, D. Reby, and N. Mathevon, "Fathers are as good as mothers at recognizing the cries of their baby," *Nature Communication*, vol. 4, no. 1, pp. 1-6, 2013.

[13] Serena Gordon. Health day. [Online]. https://consumer.healthday.com/caregiving-information-6/infant-and-child-care-health-news-410/infant-crying-can-trigger-abuse-in-some-parents-521678.html {Last accessed on May 10, 2016}

[14] P. S. Zeskind, "Infant crying and synchrony of arousal," *Evolution of Emotional Communication*, Ch. 10, January 2013.

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

[16] J. Soltis, "The signal functions of early infant crying," *J. of Behavioral and Brain Sciences*, vol. 27, no. 1, pp. 443-490, 2004.

[17] R. G. Barr, "Crying behaviour and its importance for psychosocial development in children," Encyclopedia on Early Child Development, University of British Columbia,

Canada, April 2006.

[18] A. Clarici, L. Travan, A. Accardo, U. De Vonderweid, and A. Bava, "Crying of a newborn child: alarm signal or protocommunication," *J. of Perceptual and Motor Skills*, vol. 95, no. 1, pp. 752-754, Dec. 2002.

[19] F. L. Porter, R. H. Miller, and R. E. Marshall, "Neonatal pain cries: effect of circumcision on acoustic features and perceived urgency," *J. of Child Development*, vol. 57, no. 3, pp. 790-802, Jun 1986.

[20] M. Cecchini, C. Lai, and V. Langher, "Dysphonic newborn cries akkow prediction of their perceived meaning ," *J. of Infant Behaviour and Development*, vol. 33, no. 3, pp. 314-320, Jun 2010.

[21] K. Michelson, K. Eklund, P. Leppanen, and H. Lyytinen, "Cry characteristocs of 172 healthy 1-to-7-day-old infants," *Folia Phoniatr. Logop.*, vol. 54, no. 4, pp. 190-200, Jul 2002.

[22] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. of Acous. Soc. of America*, vol. 105, no. 3, pp. 1455-1468, March 1999.

[23] J-L. Schwartz, C. Moulin-Frier, and P-Y. Oudeyer, "On the cognitive nature of speech sound systems," *J. of Phonetics*, vol. 53, no. 1, pp. 1-4, 2015.

[24] P. Lieberman, "The physiology of cry and speech in relation to linguistic behaviour," in *Infant Crying- Theoritical and Research Perspective*, Barry M. Lester and C.F. Zachariah Boukydis, Eds. New York and London: Plenum Publishing Corporation, 1985, ch. 2, pp. 29-57.

[25] H. Rothganger, "Analysis of the sounds of the child in the first year of age and a comparison to the language," *Int. J. of Early Human Development*, vol. 75, no. 1-2, pp. 55-69, Dec 2003.

[26] H. L. Golub, "A physioacoustic model of the infant cry and its use for medical diagnosis and prognosis," *J. Acous. Soc. of Amer.*, vol. 65, no. S25, pp. 825-826, 1979.

[27] H. L. Golub and M. J. Corwin, "A physioacoustic model of the infant cry," in *Infant Crying-Theoritical and Research Perspective*, Barry M. Lester and C.F. Zachariah Boukydis, Eds. New York and London: Plenum Publishing Corporation, 1985, ch. 3, pp. 59-82.

[28] National association of neonatal therapists.
[Online Available]. http://neonataltherapists.com/neonatal-auditory-development-talking-points-for-your-nicu.php {Last accessed on May 1, 2016)

[29] P. F. Ostwald, P. Peltzman, M. Greenberg, and J. Meyer, "Cries of a trisomy 13-15 infant," *Develomental Medicine and Child Neurology*, vol. 12, issue no. 4, pp. 472-477, 1970.

[30] K. Michelson, "Cry analysis of symptomless low birth weight neonates and of asphyxiated newborn infants," *Acta Paediatrica, Nurturing the Child*, vol. 60, issue, S216, pp. 9-45, Jun, 1971.

[31] H. R. Bauer and L. Zimmerman, "Newborn human cries: Prenatal cocaine exposed and nonexposed," *The J. of Acous. Soc. Of America*, vol. 95, no. 5, pp. 3013, 1994.

[32] O. Wasz-Hockert, T. Partanen, V. Vuorenkoski, and E. Valanne, "The identification of some specific meanings in the newborn and infant vocalization," *Experientia*, vol. 20, issue 3, pp. 154, 1964.

[33] K. Michelson, P. Sirvio, and O. Wasz-Hokert, "Sound spectrographic cry analysis of infants with bacterial meningitis," *Developmental Medicine and Child Neurology*, vol. 19, no. 3, pp. 309-315, 1977.

[34] K. Michelson, H. Kaskinen, R. Aulanko, and A. Rinne, "Sound spectrographic cry analysis of infants with hydrocephalus," *Acta Paediatrica Scanidinavica*, vol. 73, no. 1, pp. 65-68, 1984.

[35] M.B. Stoch and P.M. Smythe, "The effects of undernutrition during infancy on subsequent

brain growth and intellectual development," *South African Medical Journal* , vol. 41, no. 41, pp. 1027-1030, 1967.

[36] H. A. Patil, "Infant identification from their cry," in 7th *Int. Conf. on Advances in Pattern Recognition*, Kolkata, India, 2009, pp. 107-110.

[37] A. Messaoud and C. Tadj, "A cry based infant identification system," in 4th *Int. conf. on Image and Signal Process.*, 2010, pp. 192-199.

[38] Q. Xie, "Automatic infant cry analysis and recognition," The University of British Columbia, Canada), Dept. of Electrical engineering, Doctoral Thesis, 1993.

[39] Q. Xie, R. K. Ward, and Charles A. Laszlo, "Automatic assessment of infants' level of distress from the cry signals," *IEEE Trans. on Speech and Audio Process.*, vol. 4, no. 4, pp. 253-265, July 1996.

[40] Q. Xie, R. K. Ward, and C. A. Laszlo, "Determing normal infant's level of distress from cry sounds," in *Canadian Conf. on Elect. and Comp. Eng.*, vol. 2, Vancouver, BC , 2009, pp. 1094-1097.

[41] J. S. Black, "An eclectric cry reserach tool for the estimation of an infant's level of distress," University of British Columbia, Dept. of Elect. Eng., B.Sc. Thesis, 1997.

[42] S. M. Sanborn, "Effects of delayed auditory feedback on young infant's crying," University of Connecticut, Department of Arts, Masters of Arts Thesis, 2010.

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

[44] R. Nicollas, J. Giordano, and M. Ouaknine, "The very first cry: A multidisciplinary approach toward a model," *Annals of Otology, Rhinology and Laryngology*, vol. 121, no. 12, pp. 821-826, 2012.

[45] S. Rohilah, W. Mansor, L. Y. Khuan, A. Zabidi, and F. Yasmin, "An investigation into infant cry and apgar score using principal component analysis," in 5th *Int. Coll. on Sig. Proc. and its Applications*, Kuala Lumpur, 2009, pp. 209 - 214.

[46] M. Petroni, M. E. Malowany, C. C. Johnston, C. C. Johnston, and B. J. Stevens, "A new, robust vocal fundamental frequency (F_0) determination method for the analysis of infant cries," in *Proceedings of IEEE 7th Symposium on Computer-Based Medical Systems*, 1994, pp. 223-228.

[47] M. Petroni, M. E. Malowany, C.C. Johnston and B. J. Stevens "A Crosscorrelation based method for improved visualization of infant cry vocalizations," in *Canadian Conference on Electrical and Computer Engineering* , Sept. 1994, pp. 25-28.

[48] H. Fujisaki, "Automatic extraction of fundamental period of speech by autoorrelation analysis and peak detection," *Journal of Acoustic Soc. of Amer.*, vol. 32, no. 1, p. 1518, 1960.

[49] C. Manfredi, V. Tocchioni, and L. Bocchi, "A robust tool for newborn infant cry analysis," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, New York City, USA, 2006, pp. 509-512.

[50] G. Varallyay and Z. Benyo, "Melody shape - A suggested novel attribute for the biomedical analysis of the infant cry," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, Lyon, 2007, pp. 4119-4122.

[51] M. A. Ruíz, L. C. Altamirano, C. A. Reyes, and O. Herrera, "Automatic identification of qualitatives characteristics in infant cry," in *IEEE Conference on Spoken Language Technology (SLT)*, Berkeley, CA, 2010, pp. 442-447.

[52] M. A. Ruiz, C. A. Reyes, and L. C. Altamirano, "On the implementation of a method for automatic detection of infant cry units," *Procedia Eng.*, vol. 35, no. 1, pp. 217-222, 2012.

[53] J. O Garcia and C. A. Reyes Garcia, "Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, July, 2003, vo. 4, pp. 3140-3145.

[54] O. F. R. Galaviz, S. D. Cano-Ortiz, and C. A. Rayes-Garcia, "Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies," in *7th Mexican Int. Conf. on Art. Intell.*, Atizapan de Zaragoza, 2008, pp. 330-335.

[55] O. F. Reyes-Galaviz, S.D. Cano-Ortiz, and C.A. Reyes-Garcia, "Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach," in *Mexican International Conference on Computer Science (ENC)*, Baja California , 2008, pp. 261-267.

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

[57] G. Jr. Varallyay, Z. Benyo, A. Illenyi, Z. Farkas, and L. Kovacs, "Acoustic analysis of the infant cry: classical and new methods," in *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, San Francisco, CA, USA, 2004, vol. 1, pp. 313-316.

[58] G. Varallyay, "Future prospects of the application of the infant cry in medicine," *Periodica Polytechnica Ser. El. Eng.*, vol. 50, no. 1-2, pp. 47-62, 2006.

[59] J. Saraswathy, M. Hariharan, V. Vijean, S. Yaacob, and W. Khairunizam, "Performance comparison of Daubechies wavelet family in infant cry classification," in *8th International Colloquium on Signal Processing and its Applications*, Malacca, Malaysia, 2012, pp. 451-455.

[60] Y. Kheddache and C. Tadj, "Acoustic measures of the cry characteristics of healthy newborns and newborns with pathologies," *J. Biomed. Sci. and Eng.*, vol. 6, no. 8, pp. 796-804, 2013.

[61] D. Lederman et al., "Classification of cries of infants with cleft-palate using parallel hidden markov models," *Med. Biol. Eng. Comput.*, vol. 46, no. 10, pp. 965-975, 2008.

[62] D. Lederman, "Automatic Classification of Infants' Cry," University of Ben-Gurion of the Negev, Dept. of Electrical and Computer Engineering, Negev, Masters Thesis 2002.

[63] A. Zabidi, W. Mansor, Y. K. Lee, R. Sahak, and F. Y. A. Rahman, "Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism," in *5th International Colloquium on Signal Processing & Its Applications (CSPA)*, Kuala Lumpur, 2009, pp. 204-208.

[64] S. J. Sheinkopf, J. M. Verson, M. M. Rinaldi, and B. M. Lester, "Atypical cry acoustics in 6-month old infants at risk for autism spectral disorder," *J. of Autism Research*, vo. 5, no. 5, pp. 331-339, 2012.

[65] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "Classification of infant cry vocalizations using artificial neural networks," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, Detroit, MI, 1995, pp. 3475-3478.

[66] M. Petroni, M.E. Malowany, C. C. Johnston, and B. J. Stevens, "A comparison of neural network architectures for the classification of three types of infant cry vocalizations," in *IEEE 17th Annual Conference Engineering in Medicine and Biology Society* , vol. 1, Montreal, Que, 1995, pp. 821-822.

[67] S. E. Barajas-Montiel and C. A. Reyes-Garcia, "Identifying pain and hunger in infant cry with classifiers ensembles," in *International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, vol. 2, Vienna , 2005, pp. 770-775.

[68] H. E. Baeck and M. N. Souza, "Study of acoustic features of newborn cries that correlate with the context," in *Proc. of 23rd Annual Int. Conf. of IEEE, Engineering in Medicine and Biology Society (EMBS)*, Istanbul, 2001, pp. 2174-2177.

[69] R. R. Vempada, S. A. Kumar, and K. S. Rao, "Characterization of infant cries using spectral and prosodic features," in *National Conf. on Comm. (NCC)*, Kharagpur, India, 2012, pp. 1-5.

[70] A. K. Singh, J. Mukhopadhyay, and K. S. Rao, "Classification of infant cries using epoch and spectral features," in *National Conf. on Comm (NCC).*, Delhi, India, 2013, pp. 1-5.

[71] A. K. Singh, J. Mukhopadhyay, S. Kumar S.B., and K. S. Rao, "Infant cry recognition using excitation source features," in *IEEE India Conference (INDICON)*, Mumbai, India, 2013, pp. 1-5.

[72] A. Kumar Singh, J. Mukhopadhyay, and K. S. Rao, "Classification of infant cries using source, system and supra-segmental features," in *Indian Conference on Medical Informatics and Telemedicine (ICMIT)*, Kharagpur,India, 2013, pp. 58-63.

[73] J. Orozco and C. A. Reyes García, "Detecting pathologies from infant cry applying scaled conjugate gradient neural networks," in *European Symposium on Artificial Neural Network (ESANN)*, Bruges, Belgium, 2003, pp. 349-354.

[74] A. Rosales-Perez, C. A. Reyes-Garcia, J. A. Gonzalez, and E. Arch-Tirado, "Infant cry classification using genetic selection of a fuzzy model," in *17th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, vol. 7441, Argentina, 2012, pp. 212-219.

[75] M. Hariharan, J. Saraswathy, R. Sindhu, W. Khairunizam, and S. Yaacob, "Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks," in *Elsvier J. of Expert Systems with Applications*, vol. 39, no. 10, pp. 9515 - 9523, 2012.

[76] M. Hariharan, S. Yaacob, and S. Ardeen, "Pathological infant cry analysis using wavelet packet transform and probabilistic neural network," in *Elsvier J. of Expert Systems with Applications*, vol. 38, no. 12, pp. 15377-15382, 2011.

[77] R. Sahak, W. Mansor, Y. K. Lee, A. I. Mohd Yassin, and A. Zabidi, "Orthogonal least square based support vector machine for the classification of infant cry with asphyxia," in *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, Yantai, 2010, pp. 986-990.

[78] O. F. Reyes Galaviz and C. A. Reyes García, "Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system," in *Advances in Artificial Intelligence: Proc. of 4th Mexican International Conference on Artificial Intelligence (MICAI)*, Alexander Gelbukh, Alvaro Albornoz, and Hugo Terashima-Marin, Eds. Monterrey, Mexico: Springer Berlin Heidelberg, 2005, pp. 949-958.

[79] O. F. Reyes-Galaviz and C. A. Reyes-Garcia, "A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks ," in *9th Conference on Speech and Computer (SPECOM)*, St. Petersburg, Russia, 2004, pp. 1-4.

[80] D. Lederman et al., "On the use of hidden Markov models in infants' cry classification," in *22nd Convention of Electrical and Electronics Engineers*, Israel, 2002, pp. 350-352.

[81] S. D. Cano Ortiz, D. I. E. Beceiro, and T. Ekkel, "A radial basis function network oriented for infant cry classification," in *Proceedings of Progress in Pattern Recognition, Image Analysis and Applications: 9th Iberoamerican Congress on Pattern Recognition (CIARP)*, Alberto Sanfeliu et al., Eds. Puebla, Mexico: Springer Berlin Heidelberg, 2004, pp. 374--380.

[82] A. M. Goberman, S. Johnson, M. S. Cannizzaro, and M. P. Robb, "The effect of positioning on infant cries: Implications for sudden infant death syndrome," *Int. J. of Pediatric Otorhinolaryngol*, vol. 72, no. 2, pp. 153-165, Feb 2008.

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

[84] R. H. Colton, A. Steinschneider, L. Black, and J. Gleason, "The newborn infant cry: Its potential implication for development and SIDS," in *Infant Crying: Theoritical and Research Perspective*, B. M. Lester and Z. Boukydis, Eds. new York and London: Plenum Publishing Corporation, 1985, ch. 6, pp. 119-138.

[85] M. P. Robb, D. H. Crowell, P. Dunn-Rankin, and C. Tinsley, "Cry features in siblings of SIDS," *Acta Pediatrics*, vol. 96, no. 10, pp. 1404-1408, 2007.

[86] C. N. Ogbu, "Sudden infamt death syndrome (SIDS) or cot death: A review," *West Afr. J. of Med.*, vol. 22, no. 1, pp. 88-91, 2003.

[87] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "A robust and accurate cross-correlation-based fundamental frequency (F_0) determination method for the improved analysis of infant cries," in *IEEE 17th Annual Conference on engineering in Medicine and Biology Society (EMBS)*, vol. 2, Montreal, Que, 1995, pp. 975-976.

[88] M. P. Robb, D. H. Crowell, and P. Dunn-Rankin, "Sudden infant death syndrome: cry characteristics," *Int. J. of Pediatric Otorhinolaryngol.*, vol. 77, no. 8, pp. 1263-1267, Aug. 2013.

[89] ICMR, "Ethical Guidelines for Biomedical Reserach on Human Participants," New Delhi, 2006. [Online]. http://icmr.nic.in/ethical_guidelines.pdf {Last accessed on May 1, 2016}.

[90] Council for International Organizations of Medical Sciences (CIOMS), "International Ethical Guidelines for Biomedical Research Involoving Human Subjects," Geneva, 2002. [Online]. http://www.recerca.uab.es/ceeah/docs/cioms.pdf {Last accessed on May 1, 2016}.

[91] A. Chittora and H. A. Patil, "Data collection and corpus design for analysis of normal and pathological infant cry," in *Oriental COCOSDA held jointly with International Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, vol. 2, New Delhi, 2013, pp. 1-6.

[92] Civil Hospital, Ahmedabad. [Online]. http://civilhospitalamdavad.org/. {Last accessed on May 1, 2016}.

[93] N. Buddha and H. A. Patil, "Corpora for analysis of infant cry," in *Int. conf. on Speech Databases and Assessments, Oriental COCOSDA*, Hanoi, Vietnam, 2007, pp. 43-48.

[94] H. A. Patil, "Cry Baby: Using spectrographic analysis to assess neonatal health from an infant's cry," in *Advances in Speech Recognition, Mobile Environments, Call Centres and Clinics*, Amy Neustein, Ed.: Springer-Verlag, 2010, pp. 323-348.

[95] D. F. Huelke, "An overview of anatomical considerations of infants and children in the adult world of automobile safety design," in *Annual Proceedings of Association for the Advancement of Automotive Medicine*, vol. 42 , 1998, pp. 93-113.

[96] H. Arsikere, S. M. Lulich, and A. Alwan, "Estimating speaker height and subglottal resonances using MFCCs and GMMs," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 159-162, February 2014.

[97] R. D. Kent, “Instrumental assessment of children’s voice,” in *The MIT Encyclopedia of Communication Disorders, Massachusetts Institute of Technology*, pp. 36, 2004.

[98] T. F. Quatieri, *Discrete-Time Speech Signal Processing.*, issue 1, Ed. Pearson Education, 2004.

[99] C-J Thoden, A-L Jarvenpaa, and K. Michelsson, "Sound spectrographic cry analysis of pain cry in prematures," in *Infant Crying- Theoritical and Research Perspective*, Barry M. Lester and Zachariah Boukydis, Eds. New York: Plenum, 1985, vol. 1, ch. 5.

[100] D. O'Shaughnessy, "Acoustic analysis for automatic speech recognition," *Proc. of IEEE*, vol. 101, no. 5, pp. 1038-1053, 2013.

[101] A. K. Vuppala and K. S. Rao, "Speaker identification under background noise using features extracted from steady vowel regions ," *Int. J. of Adaptive Control and Signal Process.*, vol. 27, no. 1, pp. 781-792, 2013.

[102] A. Mitra, Y. V. S. Rao, and S. R. M. Prasanna, "Extraction of speaker specific excitation information from linear prediction residual of speech," *Int. J. of Speech Communication*, vol. 48, no. 10, pp. 1243-1261, 2006.

[103] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Int. J. of Speech Communication*, vol. 48, no. 10, pp. 1243-1261, 2006.

[104] K. Samudravijaya, "Hindi speech recognition," *J. of Acoustic Society of India*, vol. 29, no. 1, pp. 385-393.

[105] V. Chourasia, K. Samudravijaya, Maya Ingle, and Manohar Chandwani, "Hindi speech recognition under noisy conditions," *J. of Acoustic Society of India*, vol. 54, no. 1, pp. 41-46, Jan. 2007.

[106] R. Murali Shankar and A. G. Ramakrishnan, "Synthesis of speech with emotions," in *Int. Conf. on Comm., Computers and Devices*, Kharagpur, India, India, 2000, pp. 767-770.

[107] Z. Wu and H. Li, "Voice conversion versus speaker verification," *APSIPA Trans. on Signal and Information Process.*, vol. 3, no. 1, pp. e17-32, Jan 2014.

[108] C. d Alessandro and B. Doval, "Voice quality modification for emotional speech synthesis," in *INTERSPEECH*, Geneva, Switzerland, 2003, pp. 127-132.

[109] X. Xiao, E. S. Chng, and H. Li, "Temporal structure normalization of speech feature for robust speech recognition," *IEEE Signal Process. Letters*, vol. 14, no. 7, pp. 500 - 503, Jul. 2007.

[110] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Int. Conf. on Audio, Speech and Signal Processing(ICASSP)*, Florida, USA, 2002, pp. 1541-1544.

[111] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Int. Conf. on Speoken Language Proc. (ICSLP)*, Sydney, 1998, pp. 2819-2822.

[112] K. E. Manjunath and K. Sreenivas Rao, "Articulatory and excitation source features for speech recognition in read, extempore, and conversation modes ," *Int. J. of Speech Tech.*, vol. 18, no. 1, pp. 121-134, 2016.

[113] L. Mary and B. Yegnanarayana, "Prosodic fetaures for speaker verification," in *INTERSPEECH*, Pittsburgh, Pennsylvania, 2006, pp. 917-920.

[114] L. Mary, K. K. Anish Babu, A. Joseph, and G. M. George, "Evaluation of mimicked speech using prosodic features," in *Int. Conf. on Acous., Speech and Signal Process. (ICASSP)*, Vancouver, Canada, 2013, pp. 7189-7193.

[115] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *J. of Acoustic Society of America*, vol. 136, no. 4, pp. 1932-1941, Oct. 2014.

[116] V. K. Mittal and B. Yegnanarayana, "Study of changes in glottal vibration characteristics during laughter," in *INTERSPEECH*, Singapore, 2014, pp. 1777-1781.

[117] Factors influencing fundamental frequency. [Online Available]: http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/influence.html {Last accessed on May 15, 2016}.

[118] L. Ljung, *System Identification- Theory for the User*, 2nd ed.: Prentice Hall, 1998.

[119] B.S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The J. of Acous. Society of Amer.*, vol. 50, no. 2B, pp. 637-655, April 1971.

[120] B. S. AtaL, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154-161, 2006.

[121] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of IEEE*, vol. 63, no. 4, pp. 561-580,

1975.

[122] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals.*: Wiley-IEEE Press, 1993.

[123] H. A. Patil and T. B. Patel, "Nonlinear Prediction of Speech by Volterra-Wiener Series," in *INTERSPEECH*, Lyon, France, 2013, pp. 1687-1691.

[124] H. A. Patil, P. K. Dutta, and T. K. Basu, "On the investigation of spectral resolution problem for identification of female speakers in bengali," in *IEEE International Conference on Industrial technology (ICIT)*, vol. 1, Mumbai, 2006, pp. 375-380.

[125] H. A. Patil, "Speaker Recognition in Indian Languages: A feature based approach," Department of Electrical Engg., Indian Institute of Technology (IIT), Kharagpur, Doctoral Thesis 2005.

[126] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475-487, 1976.

[127] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599-601, Oct 1980.

[128] H. M. Teager and S.M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, William J Hardcastle and Alain Marchal, Eds. Netherlands: Springer, 1990, ch. 3, pp. 241-261.

[129] H. A. Patil and K. K. Parhi, "Development of TEO phase for speaker recognition," in *Inter. Conference on Signal Processing and Communications (SPCOM)*, Bangalore, 2010, pp. 1-5.

[130] J. F. Kaiser, "On a simple algorithm to calculate the \`energy' of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, NM, 1990, pp. 381-384.

[131] O. Wasz-Hockert, K. Michelsson, and J. Lind, "Twenty five years of Scandinavian cry research," in *Infant Crying- Theoritical and Research perspective*, Barry M. Lester and Zachariah Boukydis, Eds. New York: Plenum Publishing Corporation, 1985, ch. 4, pp. 83-101.

[132] D. Gabor, "Theory of communication," *J. of the Inst. of Radion and Comm. Eng.*, vol. 93, no. 26, pp. 429-441, November 1946.

[133] N. Samudrarajan and N. Vasudha, "Genesis of wavlet transform types and applications," in *Wavelets and Fractals in Earth System Sciences*, E. Chandrasekhar, V.P. Dimri, and V. M. Gadre, Eds.: CRC Press, Taylor and Francis Group, 2014, pp. 95-96.

[134] Medscape. [Online Available]. http://emedicine.medscape.com/article/1002527-overview {Last accessed on May 10. 2016}.

[135] J. Raes, K Michelsson, and M. Despontin, "Spectrographic analysis of the crying of infants with laryngeal disorders," *Acta Otorhinolaryngol. Belg*, vol. 34, no. 3, pp. 224-237, 1980.

[136] WebMD. [Online Available]. http://www.webmd.com/asthma/guide/what-is-asthma. {Last accessed on May 10. 2016}.

[137] Medline Plus. [Online Available]. https://www.nlm.nih.gov/medlineplus/congenitalheartdefects.html

[138] Healthline. [Online Available]. http://www.healthline.com/health/down-syndrome#Overview1. {Last accessed on May 10. 2016}.

[139] Pennstate Hershey. [Online Available] http://pennstatehershey.adam.com/content.aspx?productId=117&pid=1&gid=007301. {Last accessed on May 10. 2016}.

[140] World Health Organization. [Online Available] http://www.who.int/water_sanitation_health/diseases/malnutrition/en/. {Last accessed on May 10. 2016}.

[141] K. Juntunen, P. Sirvio, and K. Michelsson, "Cry charcateristics in infants with severe malnutrition," *European J. of Pediatrics*, vol. 128, no. 4, pp. 241-246, Dec 1978.

[142] Birth Injury Guide. [Online Available]. http://www.birthinjuryguide.org/birth-injury/types/hypoxic-ischemic-encephalopathy-hie/. {Last accessed on May 10. 2016}.

[143] NIH:National school of neurological disorders and stroke. [Online Available]. http://www.ninds.nih.gov/disorders/hydrocephalus/detail_hydrocephalus.htm. {Last accessed on May 10. 2016}.

[144] MedBroadcast. [Online Available]. http://www.medbroadcast.com/Condition/GetCondition/Meningitis. {Last accessed on May 10. 2016}.

[145] Healthline. [Online Available]. http://www.healthline.com/health/neonatal-respiratory-distress-syndrome#Overview1. {Last accessed on May 10. 2016}.

[146] Medline Plus. [Online Available]. https://www.nlm.nih.gov/medlineplus/ency/article/007322.htm. {Last accessed on May 10. 2016}.

[147] Seattle Children's Hospital Research Foundation. [Online Available]. http://www.seattlechildrens.org/medical-conditions/airway/birth-asphyxia-symptoms/.{Last accessed on May 10. 2016}.

[148] Autocorrelation Method [Online Available]. http://www.fit.vutbr.cz/~grezl/ZRE/lectures/05_pitch_en.pdf. {Last accessed on May 10. 2016}.

[149] N. A. Weiss and M. J. Hasset, *Introductory Statistics*.: Addison-Wesley, 1993.

[150] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *J. Acous. Soc. of America*, vol. 126, no. 4, pp. 2061-2071, Oct. 2009.

[151] H.A. Patil and S. Viswanath, "Effectiveness of Teager energy operator for epoch detection from speech signals," *Inter. J. of Speech Technology*, vol. 14, no. 4, pp. 321-337, 2011.

[152] A. Chittora and H. A. Patil, "Significance of unvoiced segments and fundamental frequency for infant cry analysis," in *Int. Conf. on Text, Speech and Dialogue (TSD)*, Plzen, Czech Republic, 2015, pp. 273-281.

[153] N. A. Fuamenya, M. P. Robb M P, and K. Wermke, "Noisy but effective: Crying across the first 3 months of life," *J. of Voice*, vol. 29, no. 3,pp. 281-286, 2015.

[154] "SIDS and other sleep related infant deaths: Expansion of recommendations for a safe infant sleeping environment," *Pediatrics*, vol. 128, no. 5, pp. 1341-1367, Nov. 2011. [Online]. http://pediatrics.aappublications.org/content/128/5/e1341. {Last accessed on May 10. 2016}

[155] The World Bank. [Online Available]. http://data.worldbank.org/indicator/SP.DYN.IMRT.IN. {Last accessed on May 10. 2016}.

[156] D. F. Harrison, "Histologic evaluation of the larynx in sudden infant death syndrome," *Annals of Otology, Rhinology and Laryngology*, vol. 100, no. 3, pp. 173-175, 1991.

[157] R. H. Colton, Alfred Steinschneider, Lois Black, and John Gleason, "The newborn infant cry: Its potential implication for development and SIDS," in *Infant Crying*, Barry M. Lester and C.F. Zachariah Boukydis, Eds. New York: Plenum Press, ch. 6, pp. 121-138.

[158] B. M. Lester and Z. C. F. Boukydis, Eds., *Infant Crying- Theoritical and Research perspective*.: Plenum Press.

[159] D. Bone et al., "Classifying language related developmental disorders from speech cues: The promise and the potential confounds," in *INTERSPEECH*, Lyon, France, 2013, pp. 182-186.

[160] D. Mehta and R. E. Hillman, "Use of aerodynamic measures in clinical voice assessment,"

Perspectives on Voice and Voice Disorders, vol. 17, no. 3, pp. 14-18, Nov 2007.

[161] D. D. Mehta and R. D. Hillman, "Current role of stroboscopy in laryngeal imaging," *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 20, no. 6, p. 429, Dec 2012.

[162] Sid-Ahmed Selouani, M. S. Yakoub, and D. O'Shaughnessy, "Alternative speech communication system for person with severe speech disorders," *EURASIP J. of Adv. in Sign. Process.*, vol. 1, no. 1, pp. 249-254, Jun 2009.

[163] T. Binzoni, C. S. Seelamantula, and D.Van De Ville, "A fast time-domain algorithm for the assessment of tissue blood flow in laser Doppler flowmetry," *J. of Physics in Medicine and Biology*, vol. 55, no. 13, p. N383, 2010.

[164] C. S. Seelamantula and T. V. Sreenivas, "Blocking artifacts in speech/ audio: Dynamic auditory model based characterization and optimal time-freqency smoothing," *Elsvier J. of Signal Processing*, vol. 89, no. 4, pp. 523-531, 2009.

[165] Kay Elemetrics. (1994) Massachusetts Eye & Ear Infirmary Voice Disorder Databse (MEEI Database): Elemetrics Disordered Voice Database (version 1.03).

[166] C.L. Nikias and Raghuveer R Mysore, "Bispectrum estimation: A digital signal processing framework," *Proc. of the IEEE*, vol. 75, no. 7, pp. 869-891, July 1987.

[167] C. L. Nikias and J.M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal Proc. Mag.*, vol. 10, no. 3, pp. 10-37, July 1993.

[168] Y. Haitao, Y. Wang, X. Zhanlin, and L. Wei, "Feature extraction and classification based on bispectrum for underwater targets," in *Inter. Conf. on Intelligent System Design and Engineering Application (ISDEA)*, vol. 1, Changsha , 2010, pp. 742-745.

[169] S. Li and Y. Liu, "Feature extraction of lung sounds based on bispectrum analysis," in *Third Inter. Symposium on Information Processing (ISIP)*, 2010, pp. 393-397.

[170] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253-1278, 2000.

[171] A. Chittora and H. A. Patil, "Analysis of normal and pathological infant cries using bispectrum features derived using HOSVD," in *Int. Conf. on Biosignal. Analysis, Processing and Systems (ICBAPS)*, Kuala Lumpur, 2015, pp. 151-155.

[172] A. Chittora and H. A. Patil, "Classification of normal and pathological infant cries using bispectrum features," in *23rd European Signal Processing Conf.(EUSIPCO)*, Nice, 2015, pp. 639-643.

[173] A. Chittora and H. A. Patil, "Classification of pathological infant cries using modulation spectrogram features," in *9th Int. Symp. Chinese Spoken Lang. Proc. (ISCSLP)*, Singapore, 2014, pp. 541-545.

[174] A. Chittora and H. A. Patil, "Classification of phonemes using modulation spectrogram based features for Gujarati languages," in *Int. Conf. on Asian Lang. Proc. (IALP)*, Kuching, 2014, pp. 46-49.

[175] B.W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) -Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[176] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf, "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in protein," *Bioinformatics*, vol. 21, no. 16, pp. 3369-3376, 2005.

[177] C-C Chang and C-J Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 27, no. 2, pp. 1--27, 2011.

[178] Higher order spectral analysis toolbox. [Online Available].
http://www.mathworks.nl/matlabcentral/fileexchange/loadFile.do?objectId=3013&objectTy

pe=file. {Last accessed on May 19, 2016}.

[179] NOISEX-92. [Online Available]. http://www.speech.cs.cmu.edu/comp.speech/Section/Data/noisex.html. {Last accessed on May 10. 2016}.

[180] S. Greenberg and B. E. D. Kingbury, "The modulation spectrogram in pursuit of an invariant representation of speech," in *Int. Conf. on Acous. Speech and Sig. Process. (ICASSP)*, vol. 3, Munich, 1997, pp. 1547-1550.

[181] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 1938-1948, 2011.

[182] L. Atlas and A. S. Shamma, "Joint Acoustic and Modulation frequency," *EURASIP J. on Applied signal Processing*, vol. 7, pp. 668-675, 2003.

[183] Modulation Toolbox. [Online Available]. http://www.ee.washington.edu/research/isdl/projects/modulationtoolbox. {Last accessed on May 10. 2016}.

[184] HIE details. [Online Available]. www.ncbi.nlm.nih.gov/pmc/artcles/PMC3010686. {Last accessed on May 10. 2016}.

[185] T. Raitio, A. Suni, J. Yamagishi, and H. Pulakka, "HMM based speech syntheis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 153-165, 2011.

[186] Y. Koike and J. Markel, "Application of inverse filtering for detecting laryngeal pathology," *Ann. Otoal. Rhinol. Laryngol.*, vol. 84, no. 1, pp. 117-124, 1975.

[187] P. Alku, "Glottal inverse filtering analysis of human voice production- A review of estimation and parameterization methods of glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623-650, 2011.

[188] A. Chittora and H. A. Patil, "Use of glottal inverse filtering for asthma and HIE infant cries classification," in *Int. Conf. on Asian Language Processing (IALP)*, Kuching, Sarawak, 2014, pp. 158-161.

[189] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acous. Soc. Am.*, vol. 118, no. 2, pp. 887-906, 2005.

[190] P. Ru, "Multiscale multirate spectro-temporal auditory model," University of Maryland College Park, Doctoral Thesis 2001.

[191] A. Chittora, H. A. Patil, and H. B. Sailor, "Analysis of normal and pathological infant cries using auditory spectrogram," in *Int. Conf. on Biosignal Analysis, Processing and Systems (ICBAPS)*, Kuala Lumpur, Malaysia, 2015, pp. 145-150.

[192] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of modified group delay feature in speech recognition," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 15, no. 1, pp. 190-202, January 2007.

[193] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, vol. 1, 2003, pp. 168-71.

[194] B. Yegnanarayana, D. Saikia, and T. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 3, pp. 610-623, 1984.

[195] V. Kamakshi Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentaion of continuous speech using minimum phase group delay function," *Int. J. of Speech Comm.*, vol. 42, no. 1, pp. 429-446, 2004.

[196] A. Chittora and H. A. Patil, "Modified group delay-based features for Asthma and HIE infant cries classification," in *18th Int. Conf. on Text, Speech and Dialogue (TSD), Lecture Notes in Artificial Intelligence (LNAI)*, P. Kral and V. Matousek, Eds. Switzerland: Springer Inter. Pub., 2015, pp. 595-602.

[197] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99-109, 1943.

[198] ANOVA [Online Available]. http://www.stat.cmu.edu/~hseltman/309/Book/chapter7.pdf. {Last accessed on May 10. 2016}.

[199] Calvin College. [Online Available]. http://www.calvin.edu/~scofield/courses/m145/materials/handouts/anova.pdf. {Last accessed on May 10. 2016}.

[200] Center for innovation in mathematics teaching. [Online Available]. http://www.cimt.plymouth.ac.uk/projects/mepres/alevel/fstats_ch7.pdf. {Last accessed on May 10. 2016}.

[201] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," *Proc. of 5th annual workshop on Computational Learning Theory, ACM Press*, pp. 144-152, 1992.

[202] A. D. Dileep and C. C. Sekhar, "Speaker identification using intermediate matching kernel based support vector machines," in *Forensic Speaker Recognition: Law Enforcement and Counter Terrorism*, Amy Neuitsen and Hemant A. Patil (Eds.), 2011, pp. 389-424.

[203] Statsoft. [Online Available]. http://www.statsoft.com/Textbook/Support-Vector-Machines. {Last accessed on May 10. 2016}.

[204] R. Kunwar, S. Kiran, S. Sundaram, and A. G. Ramakrishnan, "A HMM based online Tamil word recognizer," in *Tamil Internet Conference*, 2009, pp. 165-168.

[205] L. Rabiner and B H Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, pp. 4-16, Jan 1986.

[206] L. Rabiner "A tutorial on hidden Markov model and selected applications in speech recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257-286, Feb 1989.

[207] HMM [Online Available]. https://www.phonetik.uni-muenchen.de/forschung/publikationen/Schiel_HTK.txt. {Last accessed on May 10. 2016}.

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

List of Publications

Journals

1. Anshu Chittora and Hemant A. Patil, “Data Collection of Infant Cries for Research and Analysis”, in International Journal of Voice, Elsevier (in Press).
2. Anshu Chittora and Hemant A. Patil, “Spectral Analysis of Infant Cries and Adult Speech”, International Journal of Speech Technology, vol. 19, no. 4, pp. 841–856.
3. Anshu Chittora and Hemant A. Patil, “Newborn Infant’s Cry Analysis”, in International Journal of Speech Technology, vol. 19, no. 4, pp. 919-928.
4. Anshu Chittora and Hemant A. Patil, “Significance of Higher-Order Spectral Analysis in Infant Cry Classification”, under review in International Journal of Circuits, Systems & Signal Processing (CSSP) (minor revision submitted).

Book Chapters

5. Anshu Chittora and Hemant A. Patil, “Modified group delay-based features for Asthma and HIE infant cries classification,” in *18th Int. Conf. on Text, Speech and Dialogue* (*TSD*), Lecture Notes in Artificial Intelligence (LNAI), Springer, Plzen, Czech Republic, pp. 595-602 , Sept. 14–17, 2015.
6. Anshu Chittora and Hemant A. Patil, “Significance of unvoiced segments and fundamental frequency for infant cry analysis,” in *18th International Conference on Text, Speech and Dialogue* (*TSD*), Lecture Notes in Artificial Intelligence (LNAI), Springer, Plzen, Czech Republic, pp. 273-281, Sept. 14–17, 2015.

7. Kewal D. Malde, Anshu Chittora and Hemant A. Patil, "Classification of fricative using novel modulation spectrogram based features," in P. Maji et. al. (Eds.), *International Conference on Pattern Recognition and Machine Intelligence* (PReMI), Lecture Notes in Computer Science, LNCS, Springer-Verlag, Berlin Heidelberg, pp. 134-139, Germany, 2013.
8. Anshu Chittora and Hemant A. Patil, "Analysis of normal and pathological infant cry", submitted for possible publication in Hemant A. Patil (Ed.), *Voice Technologies for Reconstruction and Enhancement*, DeGruyter, New York, 2015.

Conferences

9. Anshu Chittora and Hemant A. Patil, "Classification Of Normal And Pathological Infant Cries Using Bispectrum Features," in the *23rd European Signal Processing Conference* (*EUSIPCO*) , Nice, France, pp. 639-643, 31st August - 4th September, 2015.
10. Anshu Chittora, Hemant A. Patil and Hardik B. Sailor, "Spectro-temporal analysis of HIE and asthma infant cries using auditory spectrogram," in *International Conference on BioSignal Analysis, Processing and System* (*ICBAPS*), Kuala Lumpur Malaysia, pp. 145-150, 26-28 May 2015.
11. Anshu Chittora and Hemant A. Patil, "Analysis of normal and pathological infant cries using bispectrum features derived using HOSVD," in *International Conference on BioSignal Analysis, Processing and System* (*ICBAPS*), Kuala Lumpur Malaysia, pp. 151-155, 26-28 May 2015.
12. Anshu Chittora, Hemant A. Patil and Kewal D. Malde, "Classification of stop consonants using modulation spectrogram-based features," in Proc. of 2nd *International Conference on Perception and Machine Intelligence* (*PerMin*), pp 145-150, Kolkata, Feb. 26-27, 2015.

13. Anshu Chittora, Kewal D. Malde and Hemant A. Patil, “Obstruent classification using modulation spectrogram based features,” in *17th Oriental COCOSDA Conference*, Phuket, Thailand, pp. 1-6, 10-12 September 2014.
14. Anshu Chittora and Hemant A. Patil, "Use of glottal inverse filtering for asthma and HIE infant cries classification,” in *International Conference on Asian Language Processing* (*IALP*), Kuching, Sarawak, pp. 158-161, Oct. 20-22, 2014.
15. Anshu Chittora and Hemant A. Patil, "Classification of phonemes using modulation spectrogram based features for Gujarati languages,” in *International Conference on Asian Language Processing* (*IALP*), Kuching, Sarawak, pp.46-49, Oct. 20-22, 2014.
16. Anshu Chittora and Hemant A. Patil, "Classification of pathological infant cries using modulation spectrogram features," in *9th International Symposium on Chinese Spoken Language Processing* (*ISCSLP*), pp.541-545, Singapore, 12-14, 2014.
17. Anshu Chittora and Hemant A. Patil, "Data collection and corpus design for analysis of normal and pathological infant cry," *International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE),* pp.1-6, 25-27 Nov. 2013.
18. Hemant A. Patil, Anshu Chittora and Kewal D. Malde, “Novel modulation spectrogram based features for obstruent classification,” in *International Conference on Acoustics*, New Delhi, pp.1-10, Nov. 10-15, 2013.

Brief Biodata of the Author

Anshu Chittora was born on *10*th April, *1980* in Jaipur, Rajasthan, India. She was brought up in Ajmer and completed her school education in Ajmer from Savitri Girls Senior Secondary School, Ajmer. She did her Bachelor of Engineering Degree in the discipline Electronics and Telecommunication Engineering from Mandsaur Institute of Technology (MIT), Mandsaur, Madhya Pradesh (MP), India. The institute is affiliated with Rajiv Gandhi Prodyogiki Vishvavidyalay (RGPV), Bhopal, MP, India. After completion of her Bachelor degree, she joined Poornima College of Engineering, Jaipur, Rajasthan, as teaching faculty. In *2004*, she joined Master of Technology (M. Tech.) program at Malviya National Institute of Technology (MNIT), Jaipur, Rajasthan with specialization in Electronics and Communication Engineering as a part-time candidate. Her Master's thesis was on the designing of concatenated space time code for multiple input multiple output (MIMO) systems.

In January *2011*, she joined Doctoral Program at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India. During her research work, she coauthored Journal papers, book chapters and conference papers. Her research interests are infant cry analysis and classification and speech signal processing. During her research work, she was awarded the *Best Paper Award* for her paper in the International conference on biosignal Analysis, Processing and Systems (ICBAPS' *15*), held in Kuala Lumpur, Malaysia.

She is a student member of IEEE, IEEE signal processing society (SPS) and International Speech Communication Association (ISCA).