Classification of Pathological Infant Cries and Dysarthric Severity-Level

by

Kachhi Aastha Bidhenbhai
202015003

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

MASTER OF TECHNOLOGY
in
ELECTRONICS AND COMMUNICATION

with specialization in
Wireless Communication and Embedded Systems
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

A program jointly offered with
C.R.RAO ADVANCED INSTITUTE OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

May 2022

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Electronics and Communications at Dhirubhai Ambani Institute of Information and Communication Technology & C.R.Rao Advanced Institute of Applied Mathematics, Statistics and Computer Science, and has not been submitted elsewhere for a degree,

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

Kachhi Aastha Bidhenbhai

Certificate

This is to certify that the thesis work entitled **"Classification of Pathological Infant Cries and Dysarthric Severity-Level"** has been carried out by **Kachhi Aastha Bidhenbhai** for the degree of Master of Technology in Electronics and Communications at *Dhirubhai Ambani Institute of Information and Communication Technology & C.R.Rao Advanced Institute of Applied Mathematics, Statistics and Computer Science* under our joint supervision.

Prof. (Dr.) Hemant A. Patil
Thesis Supervisor
DA-IICT Gandhinagar, India.

Dr. Hardik B. Sailor
Thesis Co-Supervisor
Ex. Samsung R&D Institute, Bangalore, India.

Acknowledgments

"In the course of life, we're mirrors, which sparkles with godly blessings. Each reflected ray carries wisdom and essence of humble gratitude inherited." Foremost, I would sincerely express my gratitude, and thanks to the almighty for this wonderful opportunity. I would sincerely like to express my gratitude to charioteers in the form of supervisors, Prof. (Dr.) Hemant A. Patil and Dr. Hardik B. Sailor for their constant and rigorous support, encouragement, and believing in me. Under their supervision, there was always a sense of surety and confidence of thesis completion with best possibel research quality. Their valuable inputs and suggestions helped me in analysing and understanding the cause and come up with the best results and proposition. I would express my deepest gratitude to my parents, grandparents, and other family for their support, motivation, affection, and trust. I would also, such as to thank PhD scholars of Speech Research Lab, Mr. Ankur T. Patil and Ms. Priyanka Gupta for sharing their valuable knowledge, time, and guidance during the thesis work.

I would also, such as to extend my deep gratitude to PRISM team at Samsung R&D Institute, Bangalore (SRI-B), India for providing me the opportunity to build my thesis on socially relevant research problem.

I would also grab the opportunity to thank my friends and co-authors, Anand Therattil, Piyush Chodingala, and Shreya Chaturvedi for their constant support and motivation. I would also, such as to thank Mr. Teja Vardhan Reddy, Mr. Lalit Sheoran, Ms. Ami Pandat, Mr. Rishabh Pandit, and Prof Pritam Anand for providing the best environment in the campus. I would also, such as to thank housekeeping staff for helping us in keeping the lab and campus clean.

Last but not the least, I would, such as to specially thank the authorities of DA-IICT Gandhinagar for providing a quality environment and optimum research facilities to carry out this masters thesis work.

Contents

Abstract **vi**

List of Principal Symbols and Acronyms **ix**

List of Tables **x**

List of Figures **xii**

1 Introduction **1**
1.1 Speech Production Mechanism 1
1.2 Speech Pathology 2
1.2.1 Infant Cry Signal 2
1.2.2 Dysarthria 4
1.3 Motivation 4
1.4 Social Relevance from this Thesis 5
1.5 Contributions of This Thesis Work 7
1.6 Organization of the Thesis 7
1.7 Chapter Summary 9

2 Literature Survey **10**
2.1 Introduction 10
2.2 Infant Cry Analysis: Recent Trends 10
2.3 Dysarthric Speech Analysis 12
2.3.1 Dysarthric Speech Databases 13
2.4 Chapter Summary 15

3 Experimental Setup **17**
3.1 Introduction 17
3.2 Database Details 17
3.2.1 Infant Cry Classification 17
3.2.2 Dysarthric Severity-Level Classification 18

3.3 Classifiers . . . 18
3.3.1 Gaussian Mixture Model (GMM) . . . 18
3.3.2 Support Vector Machine (SVM) . . . 19
3.3.3 Convolutional Neural Network (CNN) . . . 19
3.3.4 Light Convolutional Neural Network (LCNN) . . . 22
3.3.5 Residual Neural Network (ResNet) . . . 23
3.4 Performance Evaluation Metrics . . . 24
3.4.1 Confusion Matrix . . . 24
3.4.2 % Classification Accuracy . . . 24
3.4.3 % Equal Error Rate (EER) . . . 25
3.4.4 $F1-$Score . . . 25
3.4.5 J-Measure . . . 25
3.4.6 Matthew's Correlation Coefficient (MCC) . . . 26
3.4.7 Jaccard Index . . . 26
3.4.8 Hamming Loss . . . 26
3.4.9 Linear Discriminate Analysis (LDA) . . . 26
3.5 Chapter Summary . . . 27

4 CQCC for Infant Cry Classification **28**
4.1 Introduction . . . 28
4.2 Proposed Work . . . 28
4.2.1 Constant-Q Transform (CQT) . . . 28
4.2.2 Form-Invariance Property . . . 30
4.3 Experimental Results . . . 31
4.3.1 Results on Baby Chilanto Database . . . 31
4.3.2 Results on DA-IICT Database . . . 32
4.4 Narrowband Spectrogram of Infant Cry Modes . . . 33
4.5 Performance Evaluation . . . 37
4.5.1 Performance Evaluation using Violin Plots . . . 37
4.5.2 Performance Evaluation using $F1-$Score and J-Statistics . . . 37
4.5.3 Performance Evaluation using Latency Period . . . 38
4.6 Results Under Signal Degradation Conditions . . . 39
4.6.1 Results . . . 39
4.7 Chapter Summary . . . 40

5 Uncertainty Principle for Infant Cry Classification **41**
5.1 Introduction . . . 41
5.2 Time-Bandwidth Product . . . 41

5.3 Feature Vector Extraction Procedure . . . 43
5.4 Experimental Results . . . 44
5.5 Analysis of Latency Period . . . 46
5.6 Chapter Summary . . . 47

6 TECC For Speech Pathologies **48**
6.1 Introduction . . . 48
6.2 Proposed Feature Set . . . 48
6.3 Experimental Analysis . . . 49
6.3.1 Spectrographic Analysis of Infant Cry . . . 49
6.3.2 Teager Energy Operator (TEO) Profile Analysis . . . 50
6.4 Experimental Results . . . 52
6.5 Performance Evaluation . . . 54
6.5.1 Infant Cry . . . 54
6.5.2 Dysarthric Speech . . . 55
6.6 Chapter Summary . . . 56

7 Energy-Based Feature for Dysarthric Speech Analysis **57**
7.1 Introduction . . . 57
7.2 Proposed Work . . . 57
7.3 Analysis of LEO Profiles . . . 58
7.4 Experimental Results . . . 59
7.4.1 Performance Evaluation . . . 60
7.4.2 Linear Discriminate Analysis (LDA) . . . 60
7.5 Chapter Summary . . . 61

8 Summary and Conclusions **62**
8.1 Summary . . . 62
8.2 Limitations of the Current Work . . . 62
8.3 Future Research Directions . . . 63

9 List of Publications **65**

References **67**

Abstract

Vocal communication is the most important part of any individual's life to convey their needs. Right from the first cry of neonates to the matured adult speech, required proper brain co-ordination. Any kind of lack in coordination between brain and speech producing system leads to pathology. Asphyxia, asthma, Sudden Death Syndrome, Deaf (SIDS), etc. are some of teh infant cry pathologies and neuromotor speech disorders such as Dysarthria, Parkinson's Disease, Cerebal Palsy, etc. are some of the adult speech-related pathologies. These pathologies lead to damaged or paralysed articulatory movements in speech production and rendering unintelligible words. Infants as well as adults suffering from any of the pathologies face difficulties in conveying the emotions.

The infant cry classification and analysis is a highly non-invasive method for identifying the reason behind the crying. The present work in this thesis is directed towards analysing and classifying the normal *vs.* pathological cries using signal processing approaches. Various signal processing methods, such as Constant Q Transform (CQT), Heisenberg's Uncertainty Principle (U-Vector) and Teager Energy Operator (TEO) are analysed in this thesis. Spectrographic analysis using ten different cry modes in a cry signal is also analysed in this work. In addition to this, an attempt has also been made to analyse various pathologies using the form-invariance property of the CQT. In addition to the infant cry analysis, classification of normal *vs.* pathological cries using *10*-fold cross-validation on Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) have been adopted.

In recent the years, dysarthria has also become one of the major speech technology issue for models, such as Automatic Speech Recognition systems. Dysarthric severity-level classification, has gained immense attention via researchers in the recent years. The dysarthric severity-level classification aids in knowing the advancement of the disease, and it's treatment.

In this thesis, the dysarthric speech has been analysed using various signal processing operators, such as TEO, and Linear Energy Operator (LEO) for four different dysarthric severity-level against normal speech. With increasing use of

artificial intelligence, there has been a significant increase in the use of deep learning methods for pattern classification task. To that effect, the severity-level classification of dysarthric speech, deep learning techniques, such as Convolutional Neural Network (CNN), Light-CNN (LCNN), and Residual Neural Network (ResNet) have been adopted. Finally, the performance of various signal processing-based feature has been measured using various performance evaluation methods, such as $F1$-Score, J-Statics, Matthew's Correlation Coefficient (MCC), Jaccard's Index, Hamming Loss, Linear Discriminant Analysis (LDA), and latency period for the better practical deployment of the system.

Keywords: *Infant Cry Analysis, Dysarthric severity-level Classification, Constant Q Transform, Teager Energy Operator, Deep Learning*

List of Principal Symbols and Acronyms

AIDS Assessment of Intelligibility of Dysarthric Speech

ASR Automatic Speech Recognition

CNN Convolutional Neural Network

CQCC Constant Q Cepstral Coefficients

CQT Constant Q Transform

DEB Dysarthria Examination Battery

FDA Frenchay Dysarthria Assessment

GMM Gaussian Mixture Model

LCNN Light CNN

LCNN Light Convolutional Neural Network

LDA Linear Discriminant Analysis

LEO L^2 norm Energy Operator

LFCC Linear Frequency Cepstral Coefficients

MCC Matthew's Correlation Coefficients

MFCC Mel Frequency Cepstral Coefficients

ResNet Residual Neural Network

ResNet Residual Neural Network

SIDS Sudden Infant Death Syndrome

SIT Speech Intelligibility Test

SLP Speech Language Pathologist

SoTA State-of-The-Art

STFT Short-Time Fourier Transform

SVM Support Vector Machine

TBP Time-Bandwidth Product

TECC Teager Energy Cepstral Coefficients

TEO Teager Energy Operator

TFD Time-Frequency Distribution

UA Univercal Access

VC Voice Convertion

WER Word Error Rate

WFT Windowed Fourier Transform

List of Tables

2.1 Available Databases for Infant Cry Analysis and Classification. After [47]. . . . 12
2.2 Summary of Work Done in Dysarthric Speech Classification 13
2.3 Databases for Dysarthric Speech 15

3.1 Number of cries in Baby Chilanto and DA-IICT Database. After [20,25,80]. . . . 18
3.2 Severity-Level labels. After [43]. . . . 18

4.1 Results for Various f_{min} (Hz). After [4]. . . . 31
4.2 Results for Various Window Functions. After [4]. . . . 32
4.3 Results w.r.t. Various GMM Mixtures. After [4]. . . . 32
4.4 % Classification Accuracy and % EER for different Feature Sets. After [4]. . . . 32
4.5 Results for various f_{min} (Hz). After [44]. . . . 32
4.6 Results for various Window Functions. After [44]. . . . 33
4.7 Results w.r.t. Various GMM Mixtures. After [44]. . . . 33
4.8 Performance Measures for Classification Experiments on Baby Chilanto Database. After [44]. . . . 38
4.9 Results (in % Classification Accuracy) after Adding Babble Noise on Baby Chilanto Database. After [44]. . . . 40
4.10 Results (in % Classification Accuracy) after Adding Babble Noise on DA-IICT Database. After [44]. . . . 40

5.1 % Accuracy for Spectral and Cepstral u-vector. After [7]. . . . 44
5.2 % Classification Accuracy for various Cepstral and spectral Feature Set. After [7]. . . . 45
5.3 % Classification Accuracy of ω-vector with Various Number of Subband Filters. After [7]. . . . 46

6.1 Results for Various Cepstral Feature Sets. After [9]. . . . 52

6.2 Results for Various Spectral Feature Sets on GMM and SVM. After [9]. 53
6.3 Results w.r.t. Various GMM Mixtures. After [9]. 53
6.4 Results for various Subband Filters. After [9]. 53
6.5 Results for Various Classification Systems. After [6]. 53
6.6 Confusion Matrix for MFCC, LFCC, and TECC using ResNet. After [6] 55
6.7 Various Statistical Measures for MFCC, LFCC, and TECC. After [6] 55

7.1 Results For various Classification Systems. After [5]. 60
7.2 Confusion Matrix for MFCC and LECC using CNN. After [5]. . . . 60
7.3 Various Statistical Measures of MFCC and LECC. After [5]. 60

List of Figures

1.1 Cross-Sectional View of Speech Production Anatomy. After [58]. . . 2
1.2 Cross-Sectional View of Cry Production Anatomy. After [3]. 3
1.3 Schematic Flowchart of Thesis Structure. 8

4.1 Cry Modes in Asphyxia Cry using STFT. After [44]. 34
4.2 Spectrograms of Normal Infant Cries and Their Corresponding Cry Modes. After [44]. 35
4.3 Panels (I) and (II) shows the Spectrographic Analysis of Healthy (Normal) and Pathological (Asphyxia) Infant Cry Signal: (a) the Waterfall Plot for STFT, (b) the top view of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the top view of the CQT Waterfall Plot. After [4] 36
4.4 Cry Modes in Asphyxia Cry using CQT. After [44]. 36
4.5 Cry modes in Pain cry using CQT. After [44]. 37
4.6 Violin Plots for the Experiments Performed using (a) CQCC, (b) MFCC, and (c) LFCC Feature Sets. After [44]. 38
4.7 Latency Analysis of CQCC Feature Set. After [44]. 39

5.1 Functional Block Diagram of u-vector, t-vector, and ω-vector Feature Extraction. After [7]. 43
5.2 Spectrograms of (a) Healthy *vs.* (b) Pathological Cries. After [7] . . 45
5.3 Latency Period *vs.* % Accuracy Between the Various spectral Features for CQT, u-vector, t-vector, and ω-vector. After [7]. 47

6.1 Functional Block Diagram of the Proposed Subband TEO representation and TECC Feature Set. (SF: Subband Filtered Signal, TE: Teager Energies, AE: Averaged Energies over frames). After [29,51]. . . 49
6.2 Panel-I and Panel-II represents the Spectrographic Analysis for Normal *vs.* Asphyxia Cry Samples, respectively. Fig. 6.2(a), Fig. 6.2(b), and Fig. 6.2(c) represents the STFT, MelFB, and Subband-TE Representations, respectively. After [9] 50

6.3 Subband Filtered Signal for Male Speakers Around $1^{st} Formant = 500Hz$ (Panel I) and corresponding TEO Profile (Panel II) for (a) Normal, Dysarthic Speech with Severity as (b) Very Low, (c) Low, (d) Medium, and (e) High. After [67]. 51
6.4 DET Plots for Different Feature Sets using Classifiers, namely, GMM and SVM for Infant Cry Classification. After [9]. 54
6.5 Latency period vs. % Classification Accuracy comparison between MFCC, MelFB, LFCC, LinFB, TECC, and Subband-TE. Best viewed in colour. After [6]. 56

7.1 Functional Block Diagram of the Proposed TECC and LECC Feature Sets. (SF: Subband Filtered Signal, SE: Squared Linear Energies, TE: Teager Energies, AE: Averaged Energies over frames). After [29]. 58
7.2 Subband Filtered Signal (for Vowel $/e/$) for Male Speakers Around 1^{st}Formant$F_1 = 500Hz$ (Panel I), Corresponding TEO Profile (Panel II), and Corresponding $|.|^2$ Envelope (Panel III) for (a) Normal, Dysarthic Speech with Severity-Level as (b) Very Low, (c) Low, (d) Medium, and (e) High. After [67]. 59
7.3 Scatter Plots obtained using LDA for (a) MFCC and (b) LECC. After [46]. Best viewed in colour. 61

CHAPTER 1
Introduction

Communication is the integral part of life. Proper co-ordination between brain and speech producing muscles is required for producing intelligible words. However, cry is the only way of communication for infants [55]. Hence, it makes difficult for parents or guardian to differentiate between attention seek cry or any pathological cry. Similarly, lack of brain to muscle co-ordination leads to speech impairments. These impairments can be neurogenerative or neurodegenerative [56]. Dysarthria is one of the most common speech impairment. Hence, classification of healthy *vs.* pathological cries and dysarthric severity-level is a challenging task.

1.1 Speech Production Mechanism

Speech production is the process by which thoughts are translated into speech. In ordinary fluent conversation, people pronounce roughly four syllables, ten or twelve phonemes and two-to-three words out of their vocabulary (that can contain 10 to 100 thousand words) each second. As shown in fig 1.2, speech producing organs are mainly divided into three parts normally, lungs, larynx and vocal tract. Speech is produced with pulmonary pressure supplied by the lungs that generates sound by phonation (i.e. glottal airflow) through the glottis in the larynx. This airflow is modulated by the larynx through the vocal tract, such as a periodic noisy puff. The vocal tract is further divided into oral, nasal and pharynx cavities, which gives spectral shaping to the source. In addition to the colour shaping of the source by vocal tract, air pressure variation at lips also affects the travelling sound wave that is perceived by the listener. Speech sounds are further divided into three general categories namely, *periodic, noisy* and *impulsive*. Two bands of smooth muscle tissue between the front and back of larynx is termed as *vocal folds* [30] and a slit-like time varying orifice between two vocal folds is called as glottis. There are three main states of vocal folds namely, *breathing, voiced and*

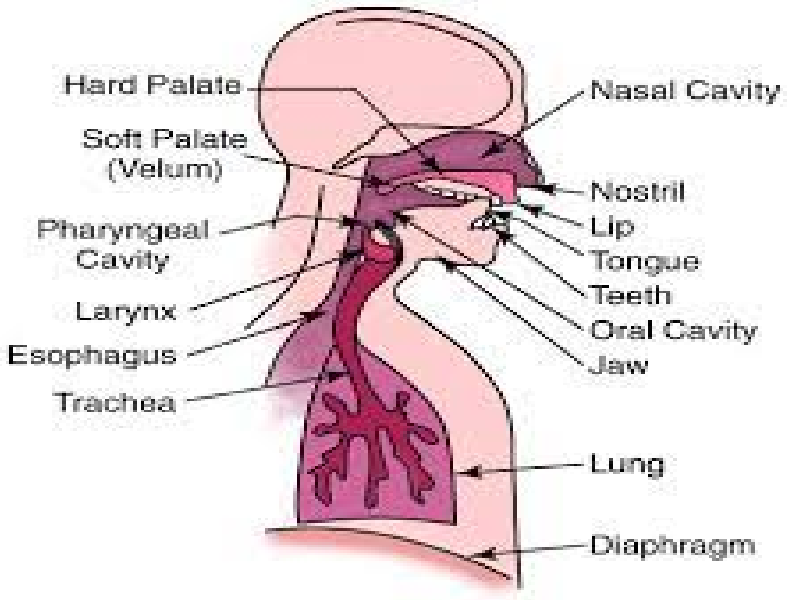


Figure 1.1: Cross-Sectional View of Speech Production Anatomy. After [58].

unvoiced. During speech production, airflow is obstructed by the vocal folds and the partial closing of glottis increases tension at the folds, which leads to self-sustained oscillations of vocal folds.

1.2 Speech Pathology

As seen in Section 1.1, lungs, vocal tract system and larynx are the main speech producing organs. Apart from these organs, to produce intelligible speech, *synchronized harmony* between several other organs, such as jaw muscle, tongue, teeth, lips, soft palate, etc is required. However, if one or more of these subsystems are abnormal or dysfunctional, the total mechanism is disrupted, rendering the output speech incoherent.

1.2.1 Infant Cry Signal

Crying is the only mode of communication for infants. However, crying requires to be a set of various complicated and sophisticated physiological activities and co-ordination between brain, vocal system, motor control mechanism and respiratory system. Crying also helps infants to develop and strengthen the pulmonary system [55]. Many a times, it is difficult for parents or guardian to determine between a normal cry or pathological cry. To overcome these situations, infant cry analysis is essential. Analytical research on infant cry analysis started as early as

1960s. Signal processing assistive tools have been made to aid parents and paediatrics in detecting the symptom of pathology and help infant get the medical aid without delay.

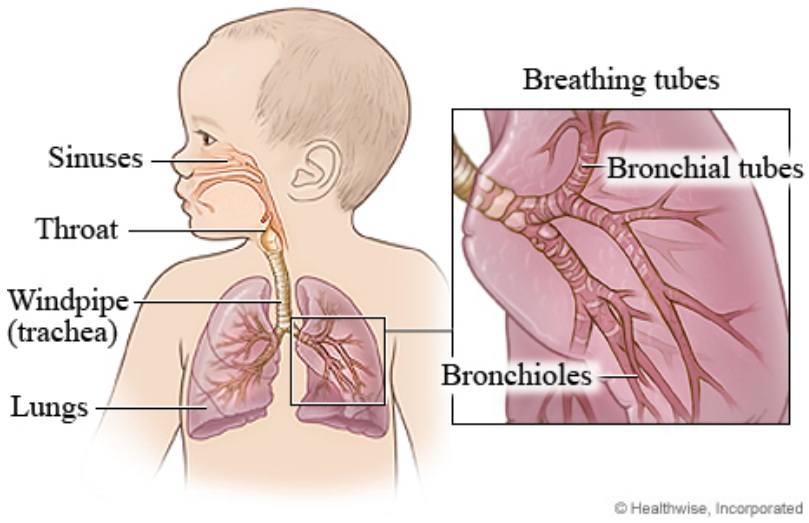


Figure 1.2: Cross-Sectional View of Cry Production Anatomy. After [3].

From the linguist view point, infant crying indicates the start of vocalization and a step towards the new language acquiring process. In the signal processing framework, it is the acoustical event, which has information regarding pitch, timbre, intonation, loudness, and rhythm. Hence, this is highly interdisciplinary in nature, where experts of different professions look at the infant cries in different perspective as a different non-verbal communication.

The foundation of infant cry analysis was laid by a Scandinavian team of researchers backs in *1960's*. The infant cry analysis proves helpful in three clinical situations, namely:

(i) Some pathologies have different characteristics, which draws attention of parents and paediatricians. Cry analysis also prevents the delayed diagnosis.

(ii) Early diagnosis of the pathology can reduce the mortality rate of infants. Infants victim of Sudden Infant Death Syndrome (SIDS) can also be prevented.

(iii) Infants on medication for abnormal cries or pathology can be monitored over the period of time to check the advancement in the treatment.

(iv) Not every infant can have luxury of being attended by trained pediatricians or pathologists more so in the developing nation context. Hence such non-invasive tool may help to generate early warning sign for any possible severe pathology.

1.2.2 Dysarthria

Dysarthria is a degenerative motor speech function impairment, which is generally the result of neurological damage in the human body. In this kind of scenarios, the person suffering from dysarthria finds difficulty in communicating and expressing vocal emotions. This impairment continues to worsen as disease progresses. Hence, its analytical study plays an important role for dysarthric severity-level diagnosis and treatment of patient. Dysarthric symptoms vary from patients-to-patients. Symptoms also depend on the impact and area of neurological impact. Some common symptoms of dysarthria are [57]:

1. Less speech loudness
2. Slow and slurry speech
3. Abnormal rhythm in speech production
4. Hoarse quality of voice
5. Articulation problem

These symptoms are common with other neurological speech disorders, such as apraxia, dysphasia, aphasia, shuttering, etc. [66]. In dysphasia and Aphasia, the person's ability to interpret and reproduce the following speech is compromised. Person suffering from apraxia, suffers from the speech planning inability, which is caused by damage in parietal lobe. On the other hand, dysarthria is caused due to incapability of one or more muscles to produce the desired speech planned by the brain.

1.3 Motivation

Around 3 million infants die within the first four months of birth due to various reasons, such as pathology, malnutrition, vaccine preventable disease, abnormalities in the brain stem controlling breathing function, etc. In the context of pathologies, birth asphyxia and related abnormalities, in particular, sudden infant death syndrome (SIDS) are the leading cause of death for infants [63]. Landmark investigations sponsored by the National Institute of Health (NIH), USA, reported evidences of abnormalities in brainstem (in particular, medulla oblongata) that is known to control breathing functions, for the infants who died of SIDS [11]. Furthermore, clinical diagnosis of asphyxia is logistics heavy and costly and thus, it is

mostly diagnosed late, however, by then, severe neurological damage would have already occurred to the infants [33]. Further, acoustic cues of the deaf infant cry depend on hearing loss, type and duration of rehabilitation and the age of pathology detection [69]. Moreover, not every infant is privileged that it is taken care by a Neonatal Intensive Care Unit (NICU) and a team of paediatricians. Analysis of infant cry signals under diverse crying reasons is necessary to avert such occurrences. Cry analysis was done via spectrographic analysis in the first two decades, where researchers employed spectrograms to define separate cry modes in the spectrogram of the infant cry. The presence of particular cry styles in babies was linked to the presence of pathology or a risk of pathology. The development of speech production throughout an individual's life starts from an infant's first babble and is transformed into fully developed speech by the age of five. For the production of speech sounds, proper coordination between the brain and the speech generating muscles is essential [56]. Lack of coordination between brain and speech producing muscles leads to speech impairments. These speech impairments, due to motor speech disorder, occurs as developmental disability. The abnormal speech creates hindrance for individuals to have effortless communication. Due to this, individual struggles to maintain the social relationship and may get prone to depression in the later age. The dysarthric severity-level classification also helps in knowing the advancement in the disease and effect of treatment on the individual. The devices must be capable of performing their intended purpose for a person with a vocal impairment, given the features of normal speech. Furthermore, due to their conventional motor impairment, people with dysarthria find it challenging to use traditional sources.

1.4 Social Relevance from this Thesis

Infant cry analysis is a noninvasive method of cry analysis that may assist doctors in the diagnosis of disease. An application designed for this purpose can be used to increase the confidence of paediatricians when making decisions about the diagnosis of specific pathologies in infants. In addition, study in this area is necessary to determine the cause of Sudden Infant Death Syndrome (SIDS), as well as to detect and diagnose newborns who are more prone to SIDS. Developing an automated cry analyser, the society can be benefited in the following ways:

i **Identifying the Needs of Infants:** Correctly identifying the cause of crying may reduce the, such aslihood of poor parenting.

ii **Developing Medical Assistive Tools:** Can help in detecting the alarming pathologies even if visual symptoms are not seen. This reduces the chances of delayed treatment. This also helps in aiding the neurological problems, where lack of medications results in medical and physical disorders.

iii **Study of Infants in Their Developmental State:** The neurological development is very fast in infants. Hence, the psychological changes related to neurological control of brain is very also very fast. These changes are also reflected in the infant cry pattern and thus, the infant cry analysis can be used to study the physical development of the infants.

iv **Reduction of Infant Mortality Rate:** The early detection of neonatal pathology may aid in lowering baby death rates.

v **Language Acquisition in Infant:** The cry patterns observed in the infant crying, depicts the way any infant acquiring the language from his or her surroundings.

vi **Speech Prosody and Therapy:** Early detection of hearing and speech problems in infants can assist parents and speech therapists in taking proactive steps toward language learning for these children. Furthermore, the crying of infants has some specific melody contour and the message delivered through the crying is in the prosodic manner.

Generally, dysarthric diagnosis requires the supervision of Speech Language Pathologist (SLP). The treatment of dysarthtic patient requires rigorous clinical treatment where patients are asked to speak various kinds of words, which includes various articulatory motion. These assessments are vulnerable to human errors. Hence, the automatic severity-level analyser method is socially relevant problem statement, which aids in knowing the advancement of treatment for dysarthric patient.

Individuals find it challenging to interpret dysarthric speech due to the character of the voice. Before someone to understand dysarthric speech, they must first become comfortable with it. Furthermore, the speaker faces some difficulties in speaking on a daily basis, which makes it a challenging duty. As a result, the Voice Conversion (VC) job is one application of dysarthric speech analysis. Dysarthric-to-normal VC can help people with dysarthria communicate more successfully with other people.

1.5 Contributions of This Thesis Work

Given the numerous challenges in infant cry research, an attempt is made to analyse and classify various types of infant cry. The areas of attention are as follows:

i **Analysis of Different Cries:** In this work, different cries are analysed. Various cry modes are analysed on the infant cries, such as hunger, pain, normal, asphyxia and deaf cries. Form-invariance structure is also analysed for various infant cries.

ii **Classification of Healthy and Pathological Cries:** Here, the healthy and pathological cries are classified. Various energy and auditory-based features were used to classify the healthy and pathological cries.

Along with given challenges faced by the researchers in the analysis of infant cries, the dysarthric severity-level classification also has several challenges. Following are the areas, which are focused on this thesis work for dysarthric severity-level classification:

i **Dysarthic Severity-Level Classification:**
Dysarthria is a neuro-motor degenerative disorder, which is caused by lack of coordination between brain and speech producing muscles. Formant frequency analysis is presented through this work.

ii **Dysarthric severity-level Classification:**
Dysarthic severity-level analysis aids in knowing the mis-cordination between the primary speech producing system or secondary speech producing system,, which aids in development of formant enhancers for automatic speech recognition. Along with this, dysarthric severity-level classification helps in knowing the advancement in the patient's condition through the treatment.

1.6 Organization of the Thesis

Figure 1.3 depicts the organization of the chapters of the thesis work as a schematic diagram, which is briefly discussed next:

- **Chapter 2** presents the detailed study on the previous investigations on infant cry analysis and dysarthric severity-level classification. Various methods based on signal processing and deep neural network aspects on various databases are also discussed.

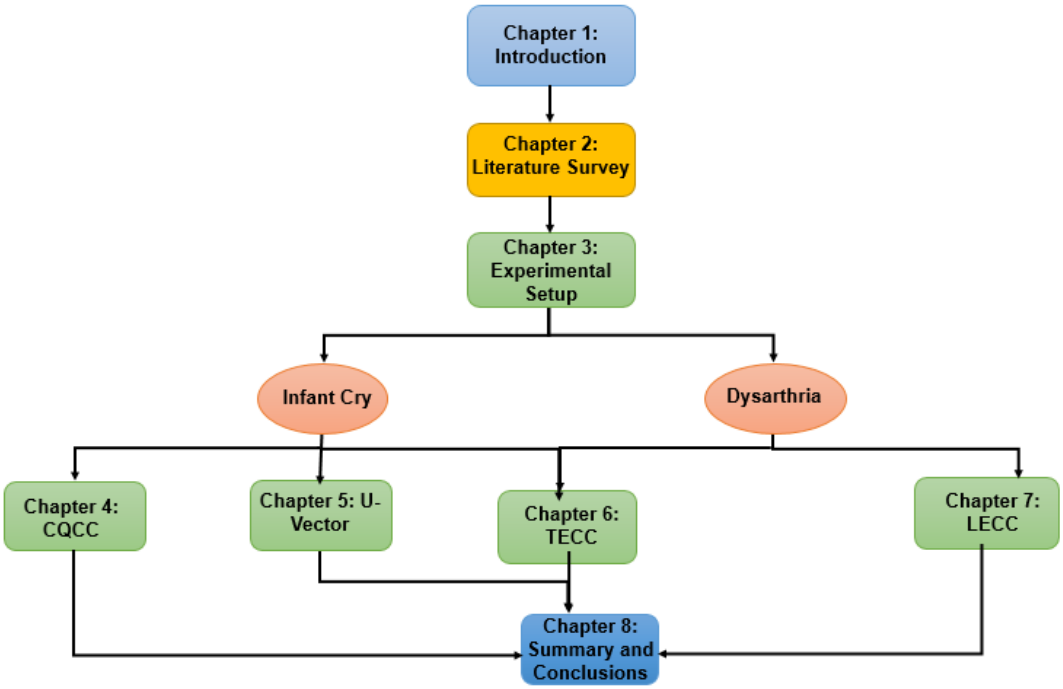


Figure 1.3: Schematic Flowchart of Thesis Structure.

- **Chapter 3** presents the details of the dataset used in this thesis work, the classifier details, and the performance measures used for comparing the models.
- **Chapter 4** presents the detailed idea and analysis based on Constant Q Transform (CQT) on infant cry. Comparison of pathological cries *w.r.t.* healthy cry is analysed. Form-invariance property of CQT, and it's significance on infant cry analysis, is also discussed in the chapter 3.
- **Chapter 5** presents the novel approach based on Heisenberg's uncertainty principle for infant cry analysis. The time-bandwidth product (TBP), u-vector
- **Chapter 6** presents the detailed investigations of dysarthric speech severity-level analysis and classification and infant cry analysis and classification using TECC feature set which is energy based signal processing feature.
- **Chapter 7** presents the extension of work on dysarthric severity-level analysis and classification using another energy-based signal processing approach using L^2 Norm based on the conclusions made in the chapter 6.
- **Chapter 8** concludes our research with an overview of the work completed within the thesis' scope. We also explain some of our work's limitations and

recommend some future research directions for improved applicability of our thesis work.

1.7 Chapter Summary

In this chapter, we looked over the brief introduction to the infant cry classification and dysarthric severity classification as a problem statement. In Section 1.1, we looked at a brief overview of how speech is produced and modulated. Next, in Section 1.2, gives the insight to the speech pathology. The motivation, social relevance and contribution from this thesis work is presented in Section 1.3, 1.4, and 1.5 respectively. Finally, Section 1.6, presents the organization and structure of the rest of the thesis. In the next chapter, we will look into the background and literature on the classification of infant cry and dysarthtic severity classification.

CHAPTER 2
Literature Survey

2.1 Introduction

In this chapter, we discuss in brief regrading the attempts that have been made in earlier times for Infant cry analysis, classification, and dysarthric severity-level classification. This chapter starts with Infant cry analysis, classification, and its recent trends. Following to this, this chapter also discuss the Dysarthric severity-level classification. This chapter also discuss the available database for both infant cry analysis and classification and dysarthric severity-level classification.

2.2 Infant Cry Analysis: Recent Trends

Foundation work of infant cry analysis started in early 1960s for four types of cries namely, pain, hunger, birth, and pleasure [88]. The detailed investigation of normal infant cry using narrowband spectrogram was pioneered by Xie. *et al.* [92]. In this work, ten different cry modes were identified, reflecting the pitch and harmonic variation in the infant cry. flat, falling, rise, vibrations, glottal roll, double harmonic breaks, dysphonation, hyperphonation, inspiration, weak vibrations. A new parameter known as the H-value was discovered using these ten distinct cry modes. This parameter can be calculated as:

$$\text{H-Value} = \frac{\text{H-type sequence number}}{\text{Total number of voiced sequences}} \tag{2.1}$$

The H-value obtained from the eq 2.1 is found to have correlation with the parents' assessment of the infant's suffering (LOD). Trailing, double harmonic breaks, dysphonation, hyperphonation, and inspiration are all examples of H-type sequences. In the Chapter 4, all the infant cry modes are analysed in different normal *vs.* pathological cries as described by Xie. *et al.*. In extension to the study presented by Xie. *et al.* on normal cries, [75] presents the study of cry

modes on pathological cries. It was observed that dysphonation and hyperphonation, are also correlated with pathological infant cries. Computer algorithms are now employed to analyse infant cry signals, enabling rapid interpretation and the development of infant cry analysis tools. Work on infant cry analysis, and classification of normal infant cries from sick infant cries, development of signal processing algorithms for infant cry analysis, and identification of cry kinds have all been done recently. The use of Mel Frequency Cepstral Coefficients (MFCC) is proposed in [65,74], for identification of an infant from his or her cry. Another interesting study on infant cry analysis shows the impact of delayed auditory feedback on crying. However, this effect does not appear to be consistent across all ages. Pitch or Fundamental frequency (F_0) may increase in some cases, however, it may also fall in others.

Attempts have also been made in the signal processing framework for infant cry signals. These techniques focus on the (F_0) estimation using auto correlation function. A new method using cross-correlation to estimate (F_0) was proposed in [76,77].

Research in classification of normal *vs.* pathological infant cry has been an emerging new research problem. The initial study on infant cry classification was done considering the normal infant cries and cries of deaf infants. The study in, the discriminative acoustic cue used for classification was the complexity of autocovariance of the cries. Following this study, the work, reported in [24], used the Support Vector Machine (SVM) for classification of normal *vs.* pathological infant cries. A novel non-invasive health care system that uses acoustic analysis of unclean noisy infant cry signals to quantitatively extract and evaluate particular cry characteristics and categorise healthy and unwell newborn newborns. The dynamic MFCC functionality uses the Gaussian Mixture Model-Universal Background Model to accomplish this (GMM-UBM). Several assistive technologies, such as baby cry analyser [1], baby pod [2], and Ubenwa mobile app [69] are developed for cost-effective and non-invasive cry diagnosis tool as a supplement to Apgar count [10] that can assist paediatrics to detect the early warning signs of various pathologies.

State-of-the-art cepstral features, such as MFCC are also used recently for cry classification task using Gaussian Mixture Model (GMM) as classifiers [8], [47]. Also, the radial based features called as Convolutional RBM are implemented in [82] using 10-fold cross-validation on traditional GMM classifiers In addition to the traditional GMM classifier, reports the infant cry classification using acoustic and prosodic features on deep neural network architectures, such as KNN, CNN,

and RNN. Finally, it was also found in, that melodic intervals in infant cry are regular phenomena in healthy infant cry. Table 2.1, shows the database available in the literature.

Table 2.1: Available Databases for Infant Cry Analysis and Classification. After [47].

Database	Creator	Recordings	Papers
Baby Chilanto	NIAOE-CONACYT, Mexico	2268	[80]
Donate a Cry	github.com/domateacry	457	[85]
Icope	infantcope.com	113	[35]
ChatterBaby	chatterbaby.org	1071	[73]
SPLANN	Hospital as part of the SPLANN study	13373	[13,87]
Self-recorded data	Recordings by 1st time parent	19691	[22]
DA-IICT	recording done by authors	1190	[20], [25]
Autism Database	recording done by authors	84	[89]
Hypothyroid Database	recording done by authors	88	[95,96]

2.3 Dysarthric Speech Analysis

The subjective assessment of dysarthric speech, which has complex characteristics requires the diagnosis assessment from Speech Language Pathologist (SLPs). SLPs mainly focus on the articulations and acoustic detection of dysarthria. There are mainly five methods in literature, which are widely used by SLPs for the assessment of dysarthria. These methods are described as follows:

- **Assessment of Intelligibility of Dysarthric Speech (AIDS):** This method takes the speaking and intelligibility rate of dysarthic speaker. However, this assessment is performed for patients above 12 years [93]. The speech of a dysarthric patient is recorded by the examiner and then played against the judges in panel, which rates the speaker on the basis of word and sentence intelligibility-level.

- **Speech Intelligibility Test (SIT):** It is the electronic form of AIDS, which was introduced in 1996 [31]. It basically provides the score of patient's speech stimuli to the examiner. The scoring process for SIT is the same as AIDS.

- **Frenchay Dysarthria Assessment (FDA):** This method determines the type of dysarthria a patient suffering from [32]. It takes various behaviours, such

as respiration, reflexes, intelligibility, movements of jaw, tongue, lip palate, etc into consideration.

- **Dysarthria Examination Battery (DEB):** This test focuses on the prosody, articulation, phonation, resonance, and respiration. The speech is rated on the scale of 5 on the basis of both word and sentence-level intelligibility. The details of DEB can be found in [72].

- **Dysarthria Profile:** This method is similar to the DEB method, where facial muscle movements are also taken into consideration. The evaluation is done by one expert clinician, one familiar and one unfamiliar listener. This provides more comprehensive assessment of dysarthric speech [79].

Table 2.2: Summary of Work Done in Dysarthric Speech Classification

Author	Feature Sets	Classifier	Classification Type	% age Accuracy
Paja et al.	Multiple acoustic feature	Mahalnobis Distance	2-level severity	95
KL Kadi et al.	Multiple acoustic feature	LDA+ GMM/SVM	4-level severity	93
Haewon Byeon et al.	Cepstral peak prominance, jitter, shimmer,etc.	Random Forest	Dysarthria vs. presbyphonia	83
C.Bhat et al.	Audio descriptor feature	ANN	4-level severity	98
Chandrashekar et al.	Mel spectrogram	CNN	3-level severity	66
M. Fernandez et al.	Log-mel spectrogram	LSTM with attention	3-level severity	77
Joshy et al.	MFCC	CNN	4-level severity	96
J.C. Vasquez-Correa et al.	STFT spectrograms	CNN	Normal vs. Dysarthria	86

2.3.1 Dysarthric Speech Databases

Standard and statistically meaningful databases play a major role in conducting research in particular reproducible research. They provide straight access to the raw material and comparative research study done by different researchers across the globe. However, the collecting the database from the dysarthric speakers is a challenge and hence, there is need for statistically meaningful database. However, there are very limited statistically meaningful database available. To that effect, it is important to discuss some statistically meaningful databases that are popularly used for dysarthric speech research. Some database details are as follows:

UA-Speech Database

It is the largest dataset available in the present time, with a total of 19 speakers, including 15 male and 4 female speakers. The age of speakers varies from $18 - 58$ years. The speaker intelligibility was rated by the naive human listener on a scale of 100 %. The recordings were done using eight microphone arrays in three blocks. There are a total of 155 repeated and common words and 100 uncommon words in each block with total of 765 word utterances [94]. The categories of the speech utterances are given below:

- 10 English digits ('Zero' to 'Nine').
- 26 radio alphabets.
- 19 computer commands.
- 100 most common words.
- 100 uncommon words from children's novels.

TORGO Database

It was developed by collaboration between department of Computer Science and Speech-Language Pathology, University of Toronto. It has been studied that Word Error Rate (WER) is of 97.5 %. Hence, this dataset is designed to develop dysarthric Automatic Speech Recognition (ASR) systems. The database consists of the data spoken by 7 different dysarthric patients (4 males, and 3 females) [81]. The database is divided into the speech samples of the following categories:

- **Non-Words** consists of 5-10 repetitions of /iy-p-ah, ah-p-iy and p-ah-t-ah-k-ah/, respectively. In addition, repeated pronunciation of high, and low pitch vowels, for 5 seconds is also included
- **Short Words** consists of English digits (1-10) with repetition and words such as yes, no, left, right, etc. In addition to this, 50 words from each word intelligibility section of FDA and 360 words from Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (YBAIDS) are also included. Ten most common words from the British National Corpus were also recorded by the subjects.
- **Restricted Sentences** consists of pre-selected phoneme rich sentences. The grandfather passage from Nemours Database [64], 162 sentences from sen-

tences intelligibility section from YBAIDS, and 460 sentences from MOCHA database.

- **Unrestricted Sentences** consists of unscripted sentences by the subjects recorded while describing 30 images of interesting situation chosen randomly from Webber Photo Cards: Story Starters Collection

HomeService Database

This dataset is designed for aiding the dysarthric patients in voice assistive applications. It contains the speech utterances of 5 dysarthric patients (3 males and 2 females) recorded through a 8 channel microphone array [68]. It consists of the two types of data as mentioned below:

- **Enrollment Data** is used to train the ASR system. The data was recorded in the closed environment for individual speaker while reading the list. n
- **Interaction Data** is captured from the consumers' homes when they control their devices. Because the identity of each word in this data is unknown, human listeners annotate it. The speech is more natural in this.

Table 2.3: Databases for Dysarthric Speech

Dataset	Speakers	Male/Females	Data	Application
TORGO Database	7	4/3	Words & Sentences	ASR
UA-Speech Database	19	15/4	Words	ASR
HomeService Database	5	3/2	Voice Commands	Voice Assistants

2.4 Chapter Summary

In this chapter, we discussed in brief regarding the previous attempts made for analysing and identifying the infant cries using various signal processing techniques and traditional classifiers, such as SVM and GMMs. We further discussed the recent trend in infant cry classification using recent Deep Learning classifiers. Table 2.1 shows the available dataset for infant cry classification purpose. Furthermore, we also discussed the methods used for dysarthric severity-level classification using Modern Artificial Intelligence and Deep learning aspects. We also

studied the the different clinical methods used by SLPs for assessment of advancement in dysarthric treatment. Further table 2.3 shows the available database for dysarthric severity-level classification.

CHAPTER 3
Experimental Setup

3.1 Introduction

This chapter discusses about the experimental setup infant cry classification and dysarthric severity-level classification. A brief description about the dataset used in this thesis work is discussed in this chapter. Furthermore, the classifiers used for classification of infant cry and dysarthric severity-level are discussed in this chapter. Finally, this chapter also discuss about various performance evaluation measures used in this thesis work for comparing the proposed feature model with the State-of-The-Art (SoTA) feature set.

3.2 Database Details

3.2.1 Infant Cry Classification

Baby Chilanto Database is used for this work. It was developed by the recordings conducted by medical doctors, which is a property of NIAOE-CONACYT, Mexico [80]. Each cry signal was segmented into one second duration (which represent one sample) and are grouped into five categories. Two groups were formed for binary classification of healthy *vs.* pathology. Healthy cry signals include three categories, namely, normal, hungry, and pain resulting in 1049 cry samples. Pathology cry signals include two categories, namely, asphyxia and deaf resulting in 1219 cry samples. Another Database used for the infant cry classification and analysis purpose is the DA-IICT Database. It was collected by [20], [25]. The sampling frequency for DA-IICT Database is 12kHz. It consists of normal and hunger cry samples for healthy infants, and asphyxia and asthma cries in pathology. Table **??** shows the statistics of Baby Chilnato and DA-IICT Database

Table 3.1: Number of cries in Baby Chilanto and DA-IICT Database. After [20, 25, 80].

Class	Category	Baby Chilanto	DA-IICT
Healthy	Normal	507	793
	Hungry	350	-
	Pain	192	-
Pathology	Asphyxia	340	215
	Deaf	879	-
	Asthma	-	182

3.2.2 Dysarthric Severity-Level Classification

In chapter 5, 6, and 7 the Universal Access Speech Corpus (UA Corpus) is used for the development of the classification system based on severity-level dysarthria. Data of 8 speakers, i.e., 4 males, namely, *M*01, *M*05, *M*07, *M*09, and 4 females, namely, *F*02, *F*03, *F*04, and *F*05 are used for this work. From 765 word utterances, 465 (155 common utterences of each blocks) utterances of microphone array number 3 per speaker are used for feature extraction as mentioned in [43]. 90% of the data was used for training and 10% for testing purpose. Each severity-level was given a different label. The label according to the severity level is mentioned as in the Table 3.2:

Table 3.2: Severity-Level labels. After [43].

Severity-Level	Label
Very Low	3
Low	2
Medium	1
High	0

3.3 Classifiers

3.3.1 Gaussian Mixture Model (GMM)

A GMM is a mixture of Gaussian probability density function (pdf) parameterized by a number of mean vectors, covariance matrices, and mixture weights of the individual mixture components. If a random vector x_n can be modeled by M Gaussian components with mean vectors μ_g, covariance matrices Σ_g, where

$g = 1, 2, ..., M$ indicate the component indices, the pdf of x_n is given by [14]:

$$f(x_n|\lambda) = \sum_{g=1}^{M} \pi_g N(x_n|\mu_g, \Sigma_g), \tag{3.1}$$

where π_g indicates the weight of the g^{th} mixture component. We denote the GMM as $\lambda = (\pi_g, \mu_g, \Sigma_g | g = 1...M)$. The likelihood of a feature vector given the GMM can be evaluated using eq. 3.1. Acoustic feature vectors in the speech literature are generally assumed to be *statistically independent*. Hence, for a sequence of feature vectors, $X = (x_n | n \in 1, ..., T)$, the probability of observing these features given the GMM is computed as [14]:

$$p(X|\lambda) = \prod_{n=1}^{T} p(x_n|\lambda). \tag{3.2}$$

A GMM is usually trained using the expectation maximization (EM) algorithm [28], which iteratively increases the likelihood between the classes.

3.3.2 Support Vector Machine (SVM)

SVM is a non-probabilistic binary linear classifier, as it assigns any new data sample directly to one of the classes. The SVM is based on discriminative training, and it gives an optimal hyperplane in the higher-dimensional feature space than the dimension of the original feature vector, given labelled training samples that categorizes new examples [14]. In particular, SVM is based on Cover's theorem on separability of patterns, which states that, the patterns that are non-linearly separable in low-dimensional feature space becomes linearly separable in high-dimensional feature space by using suitable kernel function [26].

3.3.3 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) are deep learning algorithms which uses the convolution operation in the architecture for processing the data. This convolution is done between the multidimensional input and multidimensional filter weight, known as *kernel*. The convolution operators are followed by the pooling-layer and non-linear activation operation. The combination of these three operation comprises a convolution layer, through which the features are extracted from input data. The *Fully-Connected* layers of perceptron are present in the CNN model for the classification. Further, CNN model extracts the feature similar to the

human brain by using the convolutional layers and activation functions. CNN has been widely utilized for the image classification and pattern recognition. Hence, in this study to capture the energy-based features, the CNN model has been implemented.

Convolution Operation

In CNN, convolution operations are processed by sliding the kernel through the input matrix and processing the data. The kernel size is smaller than the input matrix. The convolution operator is represented as [54]:

$$G[g_h, g_w] = (F * K)[g_h, g_w] = \sum_{i=1}^{k_h} \sum_{j=1}^{k_w} K[i,j].F[g_h - i, g_w - j], \tag{3.3}$$

where $F \in \mathbb{R}^{f_h \times f_w}$ is the input matrix, $K \in \mathbb{R}^{k_h \times k_w}$ is the kernel matrix, which is initialized randomly, and $G \in \mathbb{R}^{g_h \times g_w} \in \mathbb{R}^{f_h - k_h + 1 \times f_w - k_w + 1}$ is the output matrix. The convolution operation is performed by the elementwise multiplication of between the kernel (which slides to the next region after every operation) and the input matrix masked by the kernel. Further through the convolution operation the feature is extracted from the input matrix through the kernel, such as shapes, edges, patterns, etc.

Padding Operation

The output matrix obtained by the convolution operation has lower dimension *w.r.t* the input matrix. Hence, for deeper convolution networks, the output will diminish. Furthermore, by applying the convolution operation, it can be observed that the effects of the boundary elements are less in comparison to the elements placed at the center, which is disadvantageous if the prominent features are present at the boundaries. Hence, to overcome these disadvantages, the input matrix is padded with random values (generally zeros). Hence, the dimension of the input matrix is increased and assist in capturing the information from the boundary elements. The padding size p, for a kernel size, $k \times k$ is calculated as [54]:

$$p = \frac{k-1}{2}. \tag{3.4}$$

Stride Convolution

In the convolution operation, the kernel overlaps each element in the input matrix. However, in the larger input matrix, this represents the computational inefficiency, because the calculations are done multiple times on every element of matrix, which consumes time and memory. Further, the capturing of the global feature and local are effected through the stride values. Additionally, the stride convolution contributes in dimensionality reduction resulting in a fewer calculations, which is desirable in many cases. The output dimension, n_{out} of convolution operation implemented by padding and striding, is estimated as [54]:

$$n_{out} = \frac{n_{in} + 2.p - k}{s} + 1, \tag{3.5}$$

where n_{in} is the input matrix to the convolution layer, and k, p, and s are the kernel size, padding size, and stride length, respectively.

Activation Layers

In neural network, the output of each smallest computation unit, namely, perceptron is passed through an activation function, which introduces the non-linearities in the neural network models. Hence, the output is differentiable, which assist in the back propagation and optimization of the weights. Activation function makes the deep neural networks suitable for the complex tasks, and generalized and adaptable to the data. The activation function makes the decision of enabling the perceptron in the next layer. For the activation function $\sigma(.)$, the output z for an input, x is defined as [**?**]:

$$z = \sigma(w.x + b), \tag{3.6}$$

where w and b are the weights and bias of the perceptron, respectively. Depending on the problem, different activation selected, such as Sigmoid function, Tanh function, and Rectified Linear Unit (ReLU) function. Furthermore, the various activation functions can be used for various layers of deep neural network.

Pooling Layer

Pooling layer is utilized for the dimensionality reduction without any significant reduction in the information present. The convolution layer output is generally provided as input to the pooling layer, through which the computational complexity of the CNN reduced, making the model faster to operate. The pooling layer captures the important features and make the network less susceptible to

spatial movement from its kernel size. Therefore, the pooling layer do not affect the model performance, however, it increases the efficiency of the model.

Architecture Details

- **Dysarthria Severity-Level Classification:** In this study, CNN model was trained using Stochastic Gradient Descent (SGD) algorithm and 3 convolutional blocks each with kernel size 5×5, and 1 Fully-Connected (FC) layer [54]. The input feature is made of uniform size of $D \times 300$, where D is the dimension of the feature vector. Learning rate of 0.001 and cross-entropy loss is selected for loss estimation.

3.3.4 Light Convolutional Neural Network (LCNN)

Light-CNN (LCNN), a modified version of the neural network, has performed exceptionally for SSD task [90]. In LCNN, the non-linear activation functions are replaced with the Max-Feature-Map (MFM) activation layer, which is briefly discussed next.

Max-Feature-Map (MFM) Activation

MFM is a modified max-out function, which produces better generalization for distinct data distribution by learning with a small number of parameters. The MFM function is defined as [90]:

$$y_{ij}^{k} = max(x_{ij}^{k}, x_{ij}^{k+\frac{N}{2}}), \tag{3.7}$$

where k, i, and j represents the channel feature component, and frame number, respectively. Each convolution layer in our LCNN models applies a separate convolution operation to its input. The element-wise maximum value is selected from these two convolution layers and an output matrix is generated, which is provided as input to the next layer.

Architecture Details

- **Dysarthria Severity-Level Classification:** In this study, we utilized seven convolutional layers having MFM activation function followed by two-fully connected layers. The 1^{st} convolutional layer uses the kernel size of 5×5 and stride of, 1×1 and the following convolutional layer has a kernel size

of 3×3 and stride of 2×2 with learning rate of 0.001. Weights of the LCNN are initialized using Xavier weight initialization technique [41].

3.3.5 Residual Neural Network (ResNet)

The vanishing gradient problem in CNN introduced a new classifier, namely, ResNet, which includes the skip connections into the architecture [23].

Skip Connection

The skip connections are implemented to resolve the vanishing gradient problem of deep neural networks. The vanishing gradient occurs in several layered neural networks [23]. The gradient estimation using the back propagation is usually less than 1, which provides mode stability to the model. However, in the large networks the gradient value is very small for the initial layers, which makes the effect of initial layer insignificant. Hence, the skip layer is utilized where it passes over one or more layers in neural network layers. This provides the gradient to flow during the back propagation, such that the initial layer gradient is not 0. Further, skip connections also enable the latter layers to learn information from the initial layers. The skip connections are of two types, namely, addition and concatenation. In the addition mode, the skip connection is added to the output from the layer of the network in an elementwise manner. In concatenation mode, the output is concatenated with the skip connection and used in the densely-connected networks. This forms the residual block of the ResNet model.

Residual Blocks

The Residual blocks implemented in the ResNet model consist of two types, normal and downsampling residual blocks. The normal residual block the skip connection is connected directly with the output after skip two layers. However, in the downsampling residual block, the skip connection is connected to the output after being downsampled by the convolution layer.

Architecture Details

- **Dysarthria Severity-Level Classification:** In this study we, have utilized 12 residual blocks out of which 9 are regular residual and 3 are downsampling residual blocks. The convolution layer of 5 with stride 2 is applied along

with max pool layer of 2×2. The downsampling blocks are utilized to reduce the dimensionality of the feature maps. In the end, 1 fully connected is utilized for the multi-class classification. Similar to CNN and LCNN model, SGD with a batch size of 32 and a learning rate of 0.001 with 200 epochs.

3.4 Performance Evaluation Metrics

3.4.1 Confusion Matrix

A confusion matrix depicts, how well a classification model (or "classifier") performs on a set of test data for which the true values are known. For classification, the Confusion matrix shows how errors are distributed across the class [34]. The prediction is categorized as follows by the confusion matrix:

- **True Positive (TP)** = These are samples that are correctly predicted and belong to a specific class.
- **True Negative (TN)** = These are samples that do not belong to a specific class and are projected to belong to a different class.
- **False Positive (FP)** = These are samples that do not belong to a certain class but are projected to do so.
- **False Negative (FN)** = These are samples that are classified in one category but may be classified in another.

3.4.2 % Classification Accuracy

The major diagonal numbers of confusion matrix indicates the correct decisions made by the classifier, which gives % classification accuracy. The classification accuracy can be defined as the ratio of total number of correct decisions made by the classifier to the total number of test instances. In particular:

$$\text{Classification Accuracy (\%)} = \frac{TP + TN}{P + N}. \tag{3.8}$$

where p is total number of positive class utterances and n is total number of negative class utterances.

3.4.3 % Equal Error Rate (EER)

The EER is derived from the detection error trade-off (DET) curve, which represents the performance on detection tasks that involve the trade-off of error types [61]. In binary classification task, there are two types of errors, namely, false alarm rate ($P_{fa}(s)$) and miss rate ($P_{miss}(s)$). For arbitrary threshold s, these error rates are defined as [61]:

$$P_{fa}(s) = \frac{\text{number of pathology trials with score} > s}{\text{total number of pathology trials}}, \tag{3.9}$$

$$P_{miss}(s) = \frac{\text{number of healthy trials with score} \leq s}{\text{total number of healthy trials}}. \tag{3.10}$$

The EER refers to the threshold s_{EER} at which both the error rates are equal, i.e.,

$$EER(\%) = P_{fa}(s_{EER}) = P_{miss}(s_{EER}). \tag{3.11}$$

For *10*-fold cross-validation, a confusion matrix was combined and calculated for each fold [14].

3.4.4 $F1-$Score

Another important performance measure is $F1-$Score. $F1-$Score calculates the precision and recall for test precision is calculated by considering number of true positive results, i.e.,

$$F1 - Score = \frac{2TP}{2TP + FP + FN}. \tag{3.12}$$

The value of $F1-$Score ranges from 0 to 1. The more close value to 1, indicates perfect precision and recall of any model [34].

3.4.5 J-Measure

J-statistic ranges between -1 and 1, where -1 indicates no agreement and +1 indicated full agreement between observation and prediction. Youden's J-statistic, i.e.:

$$J - statistic = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1. \tag{3.13}$$

3.4.6 Matthew's Correlation Coefficient (MCC)

MCC indicates how closely the predicted and actual class [62] are related. When comparing models, it is usually regarded a balanced measure. MCC is in the -1 to 1 range. It is given by:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TN + FN)(TP_FN)(TN + FP)}}. \tag{3.14}$$

3.4.7 Jaccard Index

The Jaccard index compares the similarity and dissimilarity of two classes. It is worth between 0 and 1. It is given by [18]:

$$JaccardIndex = \frac{TP}{TP + FP + FN}. \tag{3.15}$$

3.4.8 Hamming Loss

It takes into account class labels that were mistakenly predicted. For prediction error (predicting an incorrect label) and missing error, all classes and test data are normalized (prediction of a relevant label). It is given by [27]:

$$\text{Hamming Loss} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j), \tag{3.16}$$

where y_i^j and $\hat{y}_i^j$ are the actual and predicted labels, and I is an indicator function. The more it is close to 0, the better is the performance of the algorithm.

3.4.9 Linear Discriminate Analysis (LDA)

LDA is primarily used for the data classification, dimensionality reduction, and data visualization, through the learning of the features, namely, Fisherfaces [46]. LDA increases the ratio of between-class variation to within-class variance in every given dataset, assuring maximum separability. Hence, through the LDA plot, the feature discriminative capabilities can be observed through the clusters formed and the distance between them.

3.5 Chapter Summary

This chapter discusses the dataset used for infant cry analysis and dysarthric severity-level classification. In addition to this, this chapter discusses the classifier details used for infant cry classification and dysarthric severity-level classification. Finally, this discusses the performance evaluation metrics used in this work for comparing the performance of various signal processing models and practical deployment of the systems. In the next chapter, we discuss about another transform called as Constant Q Transform for infant cry classification and its form-invariance property for infant cry analysis.

CHAPTER 4
CQCC for Infant Cry Classification

4.1 Introduction

In the time-frequency plane, the CQT has variable spectro-temporal resolution. CQT uses the analysis window function which is dependent on time and frequency both as parameters. Due to this, the form-invariance property is maintained in a desirable structure of feature descriptors, which is used in spectral domain for pattern classification. This structure is impossible to be maintained in traditional STFT. Adding on to it, Brown's original analysis on CQT were and inspiration for improvement of notes resolution in western music [19]. Memorizing and perception of rhythm and melody (i.e., prosody) starts around third trimester of gestation. Infants show excellent musical predisposition which has melody contour of F_0 (and it's dynamics) in most prominent form [11]. As a result, we propose to implement CQT-based feature extraction to capture melodic structure in infant cries via fundamental frequency F_0 and its harmonics, (i.e., kF_0, $k \in Z$) for infant cry classification.

4.2 Proposed Work

4.2.1 Constant-Q Transform (CQT)

In the proposed approach, we employ CQT instead of WFT in order to obtain the high frequency resolution bins in low frequency regions. Following Brown's approach [19], the frequency bins in CQT are geometrically-spaced as opposed to the linear spacing in the WFT. By selecting the appropriate parameters of the CQT, we can locate the fine structural details of spectrum of the infant cry, which are lying at very low frequency regions. Because of the geometrical spacing, the low frequency region is well emphasized. For a time-domain signal, $x(n)$, CQT maps it into the time-frequency representation such that the quality factor, Q remains con-

stant, and the center frequencies of the frequency bins are geometrically-spaced. Moreover, such constant Q analysis of the signals is desirable from both theoretical and practical viewpoints. In particular, CQT helps to preserve form-invariance property, such as the *linear* time-scaling property of the continuous-time Fourier transform which does not hold for WFT (because analysis window used in WFT is function of *only* the time parameter). Furthermore, such form-invariance property is desirable for pattern recognition applications, where we want feature descriptors of a pattern to be invariant w.r.t. scale, shift, rotation, shape, etc. [60].

Let $x(n)$ be the the speech signal then its WFT, $X(n,k)$, can be mathematically represented as [60]:

$$X(n,k) = \sum_{n=0}^{N-1} x(n) \cdot w(n,k) \cdot e^{-j(\frac{2\pi}{N})kn}. \tag{4.1}$$

where $w(n,k)$ denotes the window function, k and n represents frequency bin index and time index, respectively. The quality factor Q is defined as the ratio of the center frequency f_k of the k^{th} frequency bin to its bandwidth (Δf_k), i.e., $Q_k = f_k/\Delta f_k$. For CQT, Q_k=Q_{k-1}=Q_{k+1}=$Q \; \forall \; k \in Z$ To maintain the Q_k to be constant w.r.t. frequency, it is necessary to vary the window length in time-domain. This varying window length, $N(k)$, can be defined as: $N(k) = F_s/\Delta f_k$. Hence,

$$N(k) = \frac{F_s}{\Delta f_k} = \left(\frac{F_s}{f_k}\right) \cdot Q = T_k \cdot Q, \tag{4.2}$$

To that effect, CQT can be mathematically expressed as:

$$X^{CQT}(n,k) = \frac{1}{N(k)} \sum_{k=0}^{N(k)-1} x(n) w(n,k) e^{-j\left(\frac{2\pi}{N(k)} Qn\right)}. \tag{4.3}$$

The time-domain window $w(n,k)$ in CQT is a function of both time and frequency parameters. Hence, the resulting transform integral yields constant-Q (or constant percentage bandwidth [39]) analysis and also form-invariance property [40]. Furthermore, Discrete Cosine Transform (DCT) is applied on the CQT spectrum for feature decomposition and energy compaction to obtain the CQCC feature set.

4.2.2 Form-Invariance Property

Considering continous-time version of FT, WFT, and CQT, the time-scaling property of FT implies [38], [78]:

$$\mathcal{F}\{f(kt)\} = F_k(\omega) = \frac{1}{|k|} X\Big(\frac{k}{\alpha}\Big), \tag{4.4}$$

and a linear time-scaling of $f(t)$ corresponds to frequency scaling of $F(\omega)$ by an *inverse* factor of $\frac{1}{k}$ and vice-versa, implying the *form* or shape of spectral energy density is invariant and hence it maintains the structure. However, this is violated for the traditional WFT because the window function is dependent only on time parameter. In this context, Schroder and Atal defined WFT via practically readable (and hence, should be stable) bandpass filters [84], i.e.,

$$F_k(t,\omega) = \int_{-\infty}^{t} f(k\tau)\mathsf{w}(t-\tau)e^{-j\omega\tau}d\tau. \tag{4.5}$$

For form-invariance of WFT, we must have

$$F_k(t,\omega) = \gamma F(\alpha t, \beta\omega), k > 0 \tag{4.6}$$

where α and β are scaling factor for time and frequency, $F_k(t,\omega)$ defines WFT of $f(kt)$. However, it is shown in the literature that realization of eq. (4.6) yields the necessary and sufficient condition on weighting (i.e., window) function which belongs to the class of single term power functions, i.e., $\mathsf{w}(t) = a.t^b, t > 0$, and as per bounded input bounded output (BIBO) stability condition for Linear-Time Invariant (LTI) filter, this filter is unstable and hence, practically not realizable. However, it is interesting to note that if the window function is made to be frequency-dependent, i.e., $\mathsf{w}(t) \equiv \omega(t,\omega)$ (as in the case of CQT), in particular, equation (4.5) becomes

$$F_k(t,\omega) = \int_{-\infty}^{t} f(k\tau)\mathsf{w}(t-\tau,\omega)e^{-j\omega\tau}d\tau, \tag{4.7}$$

then the form-invariance property, i.e, eq (4.6) is satisfied by eq. (10) for the window function, i.e.,

$$\mathsf{w}(t,\omega) = v(t,\omega)t^b \qquad t > 0,\ \ \omega > 0, \tag{4.8}$$

where v(tω) is an arbitrary real function of (tω), b is real constant, and function

$\mathsf{w}(t,\omega)$ also satisfy BIBO stability condition for LTI filter [71], i.e.,

$$\int_{-\infty}^{\infty} |\mathsf{w}(t,\omega)|dt < \infty. \tag{4.9}$$

Because of this form-invariance property of CQT, we believe that it might be more suitable over spectrogram for infant cry classification task. Furthermore, equation (4.8) also holds for window function considered in most practical model and short-time analysis performed by peripheral auditory system. For example, original model developed by Flanagan based on Von Bekesy data [36] represents the window function for the mechanical spectral analysis due to the movements of basilar membrane in the cochlea of human ear [38]. In particular, $\mathsf{w}(t,\omega) = (t\omega)^2 e^{\frac{-t\omega}{2}}$, which is similar to equation (4.8).

4.3 Experimental Results

4.3.1 Results on Baby Chilanto Database

The CQT is a tune-able feature set with variable parameters such as F_{min} and window function. To focus on the desired low frequency region, the experiments initially were performed by varying F_{min}, keeping the other parameter,i.e., window function constant as Hanning window. It is observed from the Table 4.1, that f_{min}=100 Hz gives the best % classification accuracy. This might be because of the high $F0$ and the fact that its harmonics are in the range of 500 Hz and higher that it. As a result, the better discriminatory acoustic cues are captured beyond 250 Hz.

Table 4.1: Results for Various f_{min} (Hz). After [4].

f_{min}	5	10	20	50	**100**	150	200	250
Acc.	98.7	99.4	98.2	99.1	**99.8**	98.8	98.6	98.9

Further, the experiments were performed on various windowing functions keeping F_0 constant. It is observed from the Table 4.2, the optimum % classification accuracy is achieved for Hanning window. In addition, the experiments were also performed on various GMM mixtures. It is observed from Table 4.3, that the optimum results are obtained for 512 GMM mixtures.

The % classification accuracy and % EER obtained from the experiments performed using GMM and SVM as classifiers, are mentioned in Table 4.4. It is observed that the CQCC performs way better than other SoTA. Moreover the CQCC

Table 4.2: Results for Various Window Functions. After [4].

Window	Hanning	Gaussian	Hamming	Rectangular
Acc.	**99.82**	98.81	99.60	97.75

Table 4.3: Results w.r.t. Various GMM Mixtures. After [4].

Mixtures	64	128	256	512	1024
Accuracy	97.53	99.43	98.94	**99.82**	98.67

and MFCC auditory-based features uses non-linear scale (logarithmic scale in particular) along the frequency-axis. As a result, we may conclude that human auditory system-based features work better for pathological cry classification than linear-scale features.

Table 4.4: % Classification Accuracy and % EER for different Feature Sets. After [4].

		MFCC	LFCC	Cepstrals	CQCC
GMM	Acc.	98.55	98.28	98.68	**99.82**
	EER	1.23	0.50	0.47	**0.44**
SVM	Acc.	88.11	80.18	80.62	**91.19**
	EER	12.72	18.78	17.73	**6.38**

4.3.2 Results on DA-IICT Database

According to the results obtained in Section 4.3.1, the experiments were performed by varying the f_{min} keeping the constant window function as Hanning window. As mentioned in Table 4.1, it can be observed that the best possible % classification accuracy is obtained for f_{min}=100 Hz. As discussed in the previous Section, this might be due to the fact that due to high F_0 and it's harmonics (i.e., kF_0, $k \in Z$), have frequencies in the range of 500 Hz and above. Hence, the better discriminative acoustic cues are captured above 250 Hz. of frequency range. It can also be observed that initially the % classification accuracy increases as the f_{min} increases but after f_{min}=100 Hz, the % classification accuracy decreases as the f_{min} increases.

Table 4.5: Results for various f_{min} (Hz). After [44].

f_{min}	20	50	80	100	120	150
Acc.	97.62	97.03	98.42	**99.31**	97.52	98.42

Furthermore, as shown in Table 4.6, the experiments were also performed on

various window functions keeping the f_{min} constant. The best % classification accuracy was obtained for Hanning window.

Table 4.6: Results for various Window Functions. After [44].

Window	Hanning	Gaussian	Hamming	Rectangular
Acc.	**99.31**	98.61	96.44	94.44

In addition, the experiments were performed using the various number of mixtures in GMM. As observed from Table 6.3, the best result was obtained for 128 number of mixtures in GMMs. In contrast to the results obtained from Baby Chilanto Database, the best % classification accuracy was obtained for 512 mixtures in GMM. This is due to the fact that the DA-IICT database is relatively shorter, having lesser sampling frequency and lesser number of infant cry samples than the Baby Chilanto database. Thus, lesser number of mixtures in GMM are sufficient to model statistical distribution of underlying data for DA-IICT database.

Table 4.7: Results w.r.t. Various GMM Mixtures. After [44].

Mixtures	64	128	256	512	1024
Accuracy	96.53	**99.31**	98.61	96.14	94.16

4.4 Narrowband Spectrogram of Infant Cry Modes

As per the original study reported in [91], any infant cry signal can be encoded into 10 different cry modes. These cry modes have different time-frequency patterns when observed in spectrogram. These cry modes are:

(i) **Glottal Roll:** It is a gradually decreasing pattern of F_0 and total energy. It is also called as trailing cry phoneme.

(ii) **Flat:** The time-frequency pattern where, we can observe smooth and steady F_0 with less energy difference between F_0 and it's harmonics.

(iii) **Falling:** The time-frequency pattern where, we can observe descending F_0.

(iv) **Rising:** The time-frequency pattern where, we can observe ascending F_0.

(v) **Double Harmonic Break:** Weak primary simultaneous parallel harmonic series present between harmonics of F_0.

(vi) **Dysphonation:** Energy distribution is unstructured over all the frequency at either higher concentration or space between indistinguishable harmonics.

(vii) **Hyperphonation:** Energy distribution with high F_0 phonation.

(viii) **Inhalation:** It is the exhaustive expiratory phase caused by an infant's rapid breathing.

(ix) **Vibration:** Normally time-frequency pattern of high energy level with unstructured energy distribution of vibrating F_0.

(x) **Weak Vibration:** Similar to vibration with lower energy level.

For detailed analysis of class, we need to have ten distinct kinds of cry modes as shown in Figs 4.1, 4.2.

Spectrographic Analysis of Asphyxia

Asphyxia is a condition, which is caused due to lack of oxygen supply after the birth. Due to insufficient oxygen supply, the damage is done to the brain tissue. The visual symptoms of infants suffering from asphyxia are: pale skin and muscle tone. In addition, the heart rate of infants is poor.

The spectrogram shows less energy in the infant cry. Dysphonation and inhalation are prominently seen in the infant cry signal.

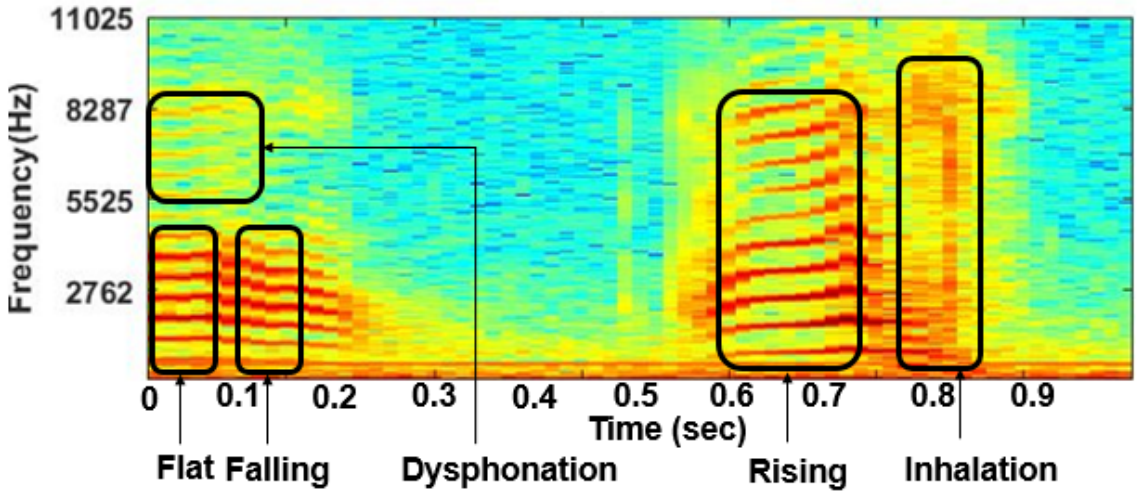
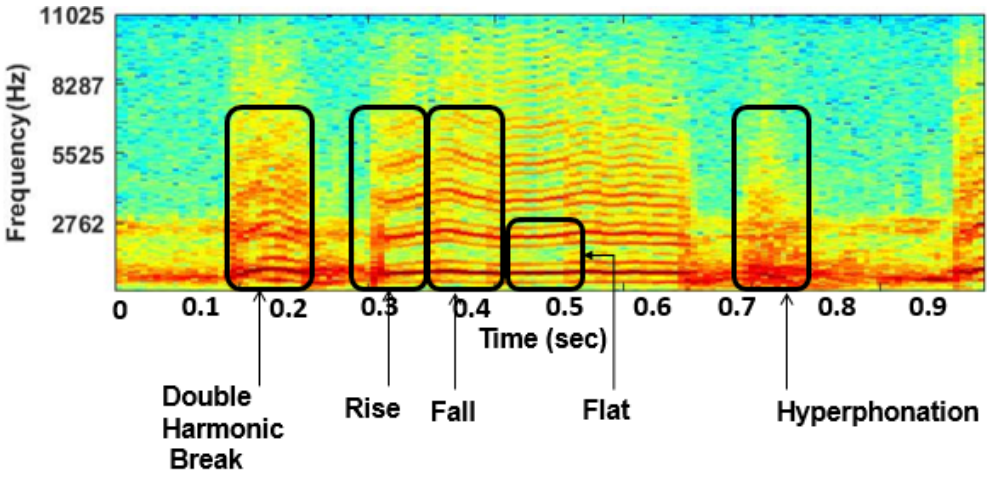


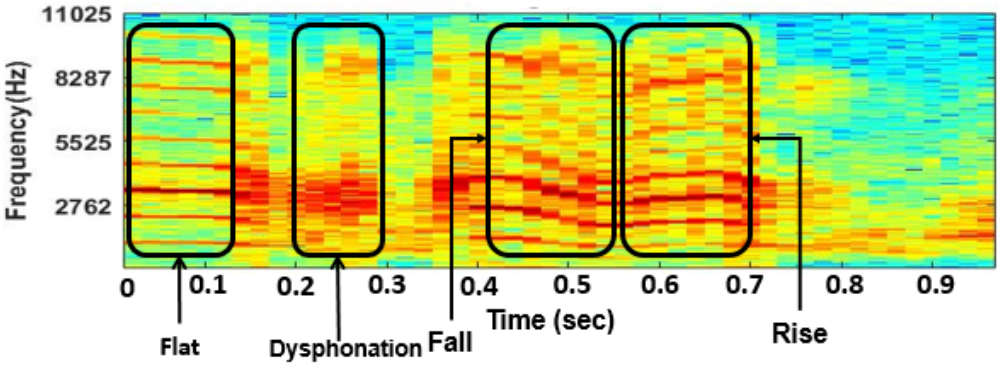
Figure 4.1: Cry Modes in Asphyxia Cry using STFT. After [44].

Spectrographic Analysis of Normal Infant Cry Signal

Figure 4.2 shows the narrowband spectrogram of normal infant cry. Melody patterns found in normal infant cry are double harmonic break, rise/fall/break (i.e., rise in start, fall in between and then flat) and hyperphonation. Apart from these, dysphonation is also found as the melody pattern in normal cry.



(a)



(b)

Figure 4.2: Spectrograms of Normal Infant Cries and Their Corresponding Cry Modes. After [44].

Spectrogram *vs.* CQTgram

In Panels I and II, the waterfall plots and corresponding top views of STFT and CQT for healthy *vs.* pathological cry signal are shown, respectively. In Fig. 4.3(a) and Fig. 4.3(b), the waterfall plot of STFT and its top view, it can be seen that F_0 of the normal signal occurs above 300 Hz. Lower frequency areas are used to estimate the abnormality in the cry signal, which appears as F_0 in pathological cry signals. CQT emphasises this anomaly significantly more due to its high frequency resolution at lower frequencies, as shown in Fig. 4.3(c) and Fig. 4.3(d).

Form-Invariance Using CQTgram

It can be observed from CQT-gram that there is a invariant structures (i.e., a pattern) in spectral energy density for CQT-gram than that of spectrogram (which has several cry modes than an invariant structure).

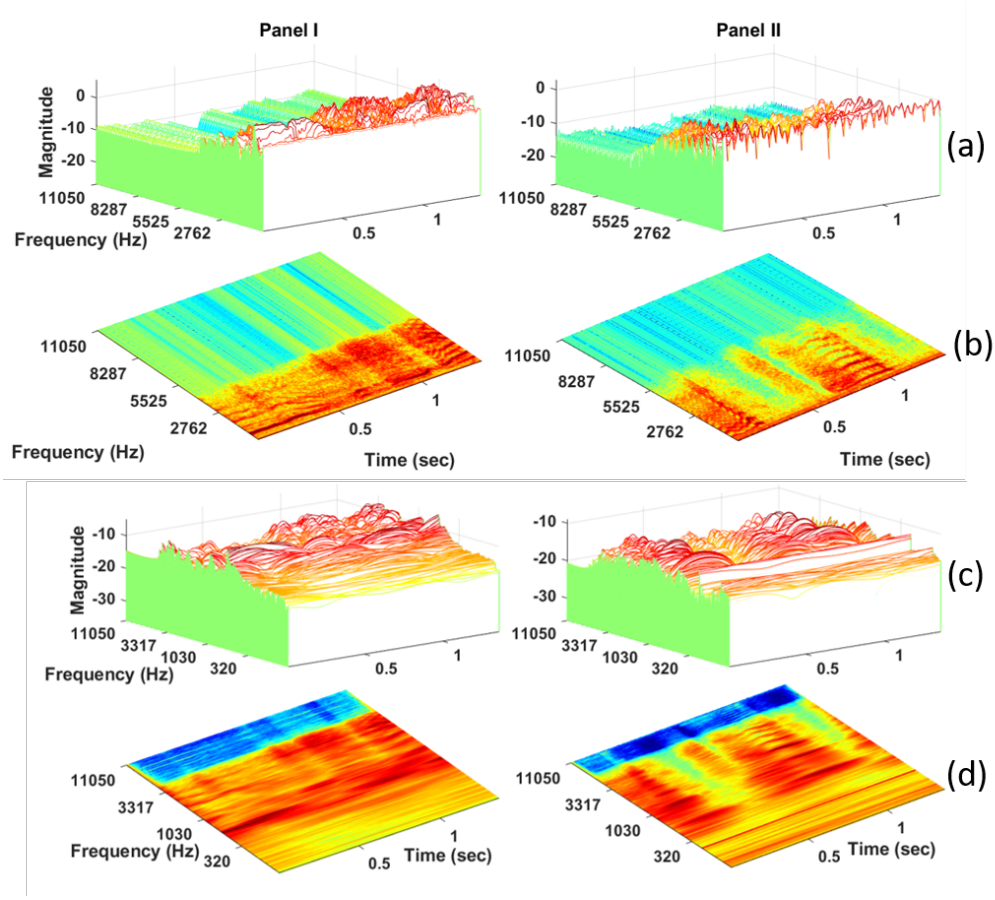


Figure 4.3: Panels (I) and (II) shows the Spectrographic Analysis of Healthy (Normal) and Pathological (Asphyxia) Infant Cry Signal: (a) the Waterfall Plot for STFT, (b) the top view of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the top view of the CQT Waterfall Plot. After [4]

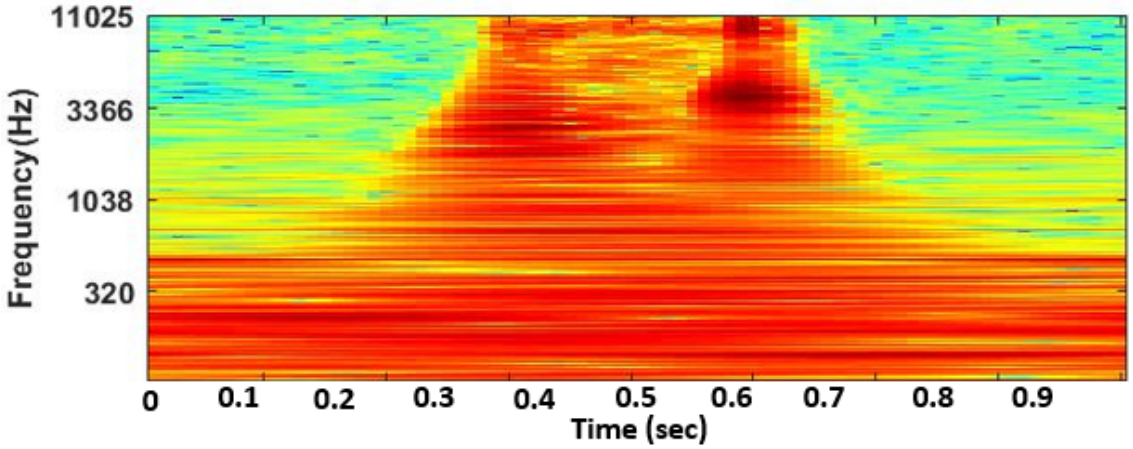


Figure 4.4: Cry Modes in Asphyxia Cry using CQT. After [44].

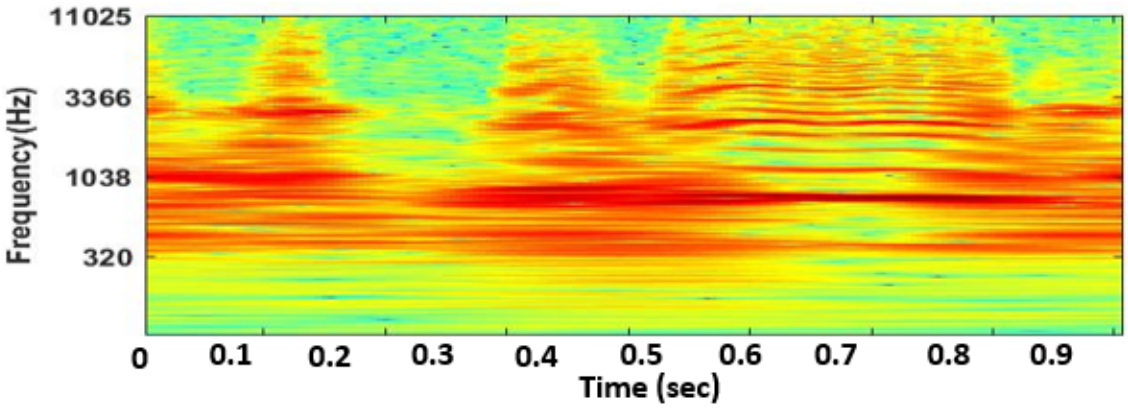


Figure 4.5: Cry modes in Pain cry using CQT. After [44].

4.5 Performance Evaluation

4.5.1 Performance Evaluation using Violin Plots

Considering the small size of the database, statistical testing is performed for the results obtained using *10*-fold cross-validation. We employed violin plots to visualize the distribution of the accuracy values obtained using *10*-fold cross-validation. Violin plot is a method for graphically demonstrating the locality, spread, and skewness groups of numerical data through their quartiles along with the addition of a rotated kernel density plot on each side, usually smoothed by a kernel density estimator. It includes a marker for the median of the data; a marker indicating the interquartile range; and possibly all sample points, if the number of samples is not too high. We have performed the *10*-fold cross-validation experiment for 50 times for each feature set and produced the **violin plots** as shown in Fig. 4.6. It can be observed from Fig. 4.6 that the mean and median values of % classification accuracy for CQCC feature set are better than the MFCC and LFCC feature sets, indicating statistical significance of the proposed CQCC feature set.

4.5.2 Performance Evaluation using $F1-$Score and J-Statistics

The % classification accuracies obtained for Baby Chilanto database is mentioned in Table 4.8. The proposed CQCC feature set obtains higher % classification accuracy compared to the MFCC, LFCC and Cepstrals on both GMM and SVM. The performance measures of the classification experiments on Baby Chilanto database is shown in Table 4.8. It can be observed that CQCC gives higher values for both the measures compared to other state-of-the-art feature sets. The

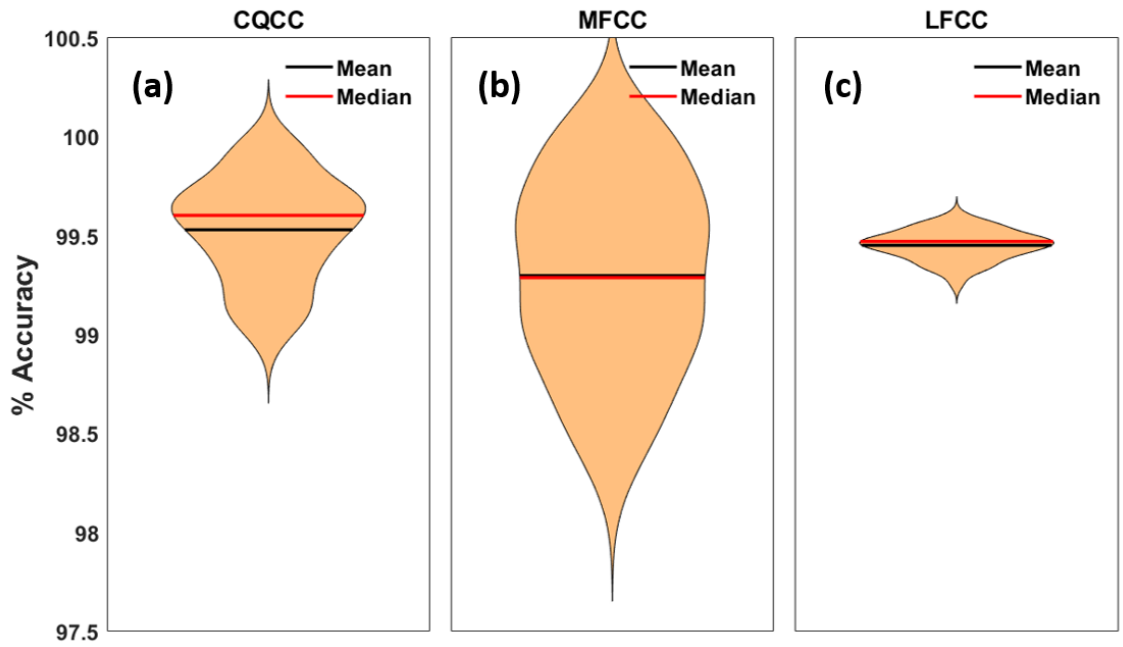


Figure 4.6: Violin Plots for the Experiments Performed using (a) CQCC, (b) MFCC, and (c) LFCC Feature Sets. After [44].

$F1-$Score takes only true positives into consideration, whereas J-statistic takes true positive as well as true negative into consideration. Hence, these measures are more meaningful evaluation parameters than % classification accuracy. Hence, the CQCC has better discriminative power for healthy *vs.* pathological infant cries for Baby Chilanto database compared to other state-of-the-art features.

Feature set	$F1-$Score	J-statistics
MFCC	0.8393	0.9801
LFCC	0.8320	0.9720
Cepstrals	0.8175	0.9622
CQCC	0.8436	0.9850

Table 4.8: Performance Measures for Classification Experiments on Baby Chilanto Database. After [44].

4.5.3 Performance Evaluation using Latency Period

Latency period analysis of the trained GMMs using state-of-the-art MFCC, LFCC, and CQCC is shown in Figure 4.7. The latency period of the trained model is estimated by computing the % classification accuracy *w.r.t.* varying durations of test speech segment in a test utterance. Latency period was analysed for cry seg-

ment of, varying from 2 ms to 20 ms. It was observed that the CQCC produces significant % classification accuracy for short duration of utterances, which are less than 6 ms. Whereas, LFCC and MFCC feature sets shows the comparable performance improvement for relatively longer cry segments of 9 ms and 20 ms, respectively. Hence, this analysis signifies suitability of the proposed CQCC feature set for practical deployment of infant cry classification system.

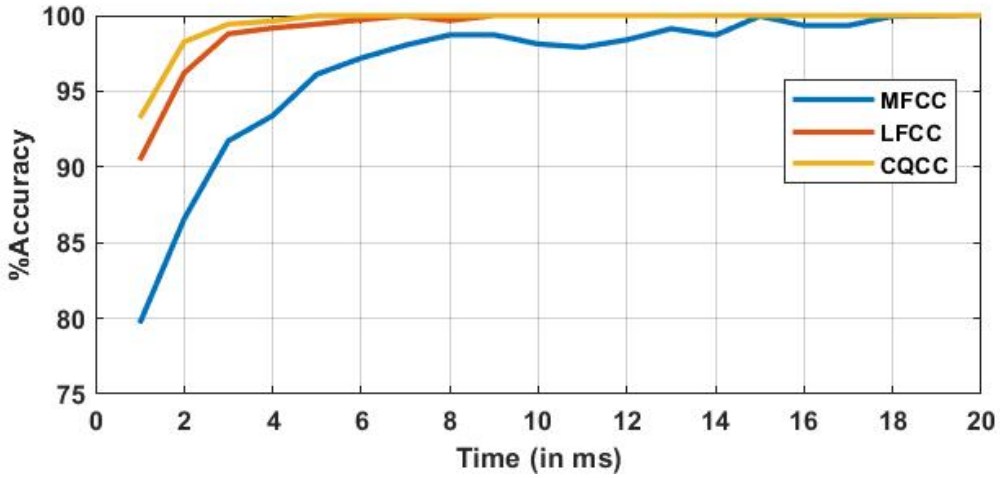


Figure 4.7: Latency Analysis of CQCC Feature Set. After [44].

4.6 Results Under Signal Degradation Conditions

4.6.1 Results

Results obtained in % classification accuracy under signal degradation conditions are reported in Table 4.9 and Table 4.10 on Baby Chilanto and DA-IICT database, respectively. It can be observed that the proposed CQCC feature set shows robust performance under signal degradation conditions with varying levels of Signal-to-Noise Ratio (SNR). It can be observed that CQCC features are relatively more robust, more so in severe signal degradation conditions (such as SNR = -5 dB and -10 dB), where MFCCs are known to get notoriously affected for several speech technology applications [42]. On the contrary, CQCC being derived from CQT having constant quality factor (Q) on entire frequency region, helps feature representation to preserve key spectral characteristics (such as resonances) of infant cry signals that are buried in noise.

Table 4.9: Results (in % Classification Accuracy) after Adding Babble Noise on Baby Chilanto Database. After [44].

SNR (in dB)	20	15	10	0	-5	-10
CQCC	**99.34**	**99.21**	**99.12**	**99.34**	**99.07**	**98.81**
MFCC	99.12	99.07	99.03	98.77	98.63	98.28
LFCC	98.90	98.99	98.77	98.68	98.59	98.50
Cepstrals	99.43	99.21	98.99	98.99	99.07	98.90

Table 4.10: Results (in % Classification Accuracy) after Adding Babble Noise on DA-IICT Database. After [44].

SNR (in dB)	20	15	10	0	-5	-10
CQCC	**97.65**	**97.65**	**97.94**	**97.84**	**97.45**	**97.35**
MFCC	97.35	96.67	96.37	95.78	94.41	93.63
LFCC	97.65	97.65	97.65	97.45	97.06	96.37
Cepstrals	97.55	97.55	97.75	97.94	97.06	97.06

4.7 Chapter Summary

In this chapter, we proposed the novel approach for capturing the melody in the infant cry classification on Baby Chilanto database and DA-IICT database. The results obtained for the proposed feature set are compared against various state-of-the-art features, such as MFCC, LFCC, and Cepstrals. *10*-fold cross-validation strategy was used to validate the experiments. The spectrographic analysis of various pathological cries against normal cry is also analysed in this chapter. Different cry modes were also analysed in this chapter. The fact that CQT follows the form-invariance property is supported by the CQTgram. The discriminative power of CQT was analysed using *10*-fold cross-validation strategy on various noise, such as babble noise and car noise. Finally, the performance of CQT was statistically measured using statistically significant performance evaluation measures, such as violin plot, $F1$-Score, J-measure, and latency period analysis on Baby Chilanto database

CHAPTER 5
Uncertainty Principle for Infant Cry Classification

5.1 Introduction

A new method using variations in both time and frequency-domains simultaneously obtained using Time-Frequency Distribution (TFD) is analysed in this chapter. TFD indicates the energy spectral density of a signal in both time and frequency domains, which represents information in the form of Heisenberg boxes. The area of the Heisenberg's box is dictated by the Heisenberg's uncertainty principle in signal processing framework [60]. The work presented in this chapter was primarily developed based on success of u-Vector for replay spoof speech detection task [45].

5.2 Time-Bandwidth Product

Let $s(t)$ be a practical non-stationary signal having Fourier transform $S(\omega) = \mathcal{F}\{s(t)\}$. If $s(t)$ has *regular* time variations, then $S(\omega)$ decays fast in high frequency region. This leads to a longer spread of energy of $s(t)$ in the time-domain [60]. In this context, from Mallat's proposition in [60] (chapter 2, proposition 2.1), a function $s(t)$ is bounded and i times continuously differentiable with bounded derivatives.

$$\int_{-\infty}^{\infty} |S(\omega)|(1+|\omega|^i)d\omega < +\infty. \tag{5.1}$$

Here, $S(\omega) = F\{s(t)\} \in L^1(R)$. However, time spread can be restricted by doing the following operation given by:

$$s_\alpha(t) = s\left(\frac{t}{\alpha}\right), \tag{5.2}$$

where the scaling factor is $\alpha < 1$. Using the time-scaling property [70], the Fourier transform of the signal $s(t)$ is

$$S_\alpha(\omega) = |\alpha| S(\alpha\,\omega). \tag{5.3}$$

Since the scaling factor $\alpha < 1$, eq. (5.3) the Fourier transform expands by a factor of $\frac{1}{\alpha}$. Thus, it shows that gain in time localization counter-affects the gain in localization in the frequency-domain and vice-versa [70].

From Heisenberg's uncertainty principle in quantum mechanics, it is impossible to find the precise location and momentum of any particle simultaneously [21]. Similarly, in signal processing framework, the energy spread in time and frequency-domain is restricted by Heisenberg's uncertainty principle [17]. Hence, the average location of the signal $s(t) \in L^2(R)$ is given as:

$$\bar{t} = \int_{-\infty}^{\infty} \frac{1}{||s||^2} t|s(t)|^2 dt, \tag{5.4}$$

and the average momentum is given by

$$\bar{\omega} = \int_{-\infty}^{\infty} \frac{1}{2\pi||s||^2} \omega|S(\omega)|^2 d\omega. \tag{5.5}$$

The obtained $\bar{\omega}$ from eq. (5.5) is also called as *effective bandwidth* by Gabor [37]. The variance, i.e., σ_t^2 and σ_ω^2 around these averages represents the uncertainty in determining the particle's position and momentum, respectively [37]. The average time variance can be calculated as

$$\sigma_t^2 = \int_{-\infty}^{\infty} \frac{1}{||s||^2} (t - \bar{t})^2 |s(t)|^2 dt, \tag{5.6}$$

and the average momentum is given by

$$\sigma_\omega^2 = \int_{-\infty}^{\infty} \frac{1}{2\pi||s||^2} (\omega - \bar{\omega})^2 |S(\omega)|^2 d\omega. \tag{5.7}$$

It can be seen from eq. (5.6) and eq. (5.7), that signal expansion in one domain results in signal contraction in the other domain. As a result, the signal spread in either domain has an inverse relationship. Hence, the Time-Bandwidth Product (TBP) given by $\sigma_t^2 \sigma_\omega^2$ is constant and represents the area of the Heisenberg's box. This product gives the "richness" of information from the infant cry segment under consideration [12, 15, 16, 53]. Given that melody contours of F_0 are prominent in children and infant cries [11], and the F_0 contours are not as *rhythmic*, and are

smeared in pathological cries. In this study, we extract discriminative features for infant cry classification using σ_t^2, σ_ω^2 and the product $\sigma_t^2\sigma_\omega^2$. To that effect, we propose t-vector, ω-vector, and u-vector features for the detection of pathological cries.

5.3 Feature Vector Extraction Procedure

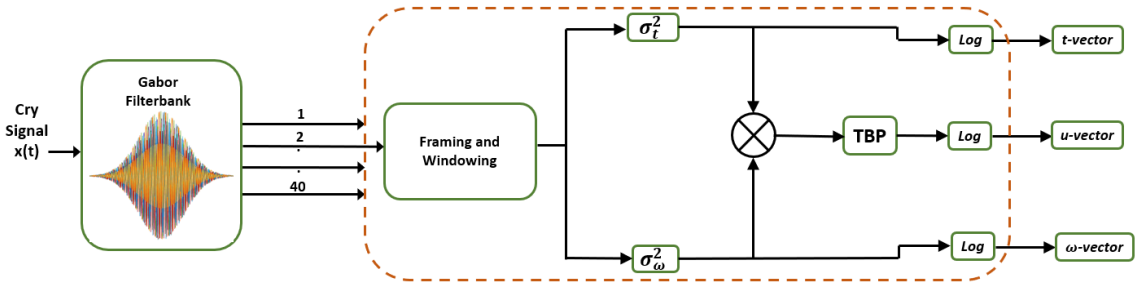


Figure 5.1: Functional Block Diagram of u-vector, t-vector, and ω-vector Feature Extraction. After [7].

The proposed feature extraction for infant cry classification is based on the fact that the spectral energy density patterns are different for healthy *vs.* pathological infant cries. This is also shown by the spectrographic analysis in Figure 5.2, which shows that pathological infant cries have high frequency of inhalation, indicating problem while breathing. Hence, spectral smearing is found in the entire frequency range. It can also be observed that there is a sudden rise in the pitch source harmonics and spreading in some regions. Therefore, the frequency variance helps to capture the regions of spectral smearing.

The feature extraction procedure in this work begins by passing the infant cry signal $s(t)$ through a Gabor filterbank of 40 subband filters. This results in 40 subband signals $s_i(t)$, where $i \in$ [1,40]. Since the cry signal is multi-component, the subband signals help in capturing frequency variances effectively [83].

Here, 40 linearly-spaced Gabor filterbank is used because of its optimal time and frequency resolution [59, 78]. Each of the subband output signals is frame blocked with a window size of 30 ms and window shift duration of 15 ms (experimentally optimized *w.r.t.* performance). For each of these frames, both σ_t^2 and σ_ω^2 is computed using the eq. (5.6) and eq. (5.7), respectively and hence, three different vector representation of the input cry signal are obtained as shown in Algorithm 1. Next, logarithmic operation is then performed on σ_t^2 and σ_ω^2 to give *t-vector* and ω *-vector* of the cry signal. Similarly, the logarithm of the product σ_t^2 σ_ω^2 gives the *u-vector* or the uncertainty vector of the cry signal, as indicated by eq.

(5.8) and eq. (5.9).

$$log(\sigma_t^2 \sigma_\omega^2) = log(\sigma_t^2) + log(\sigma_\omega^2), \tag{5.8}$$

$$u\text{-}vector = t\text{-}vector + \omega\text{-}vector. \tag{5.9}$$

Algorithm 1: TBP Computation for Infant Cry. After [7].

Input: Input: cry signal x
Output: Output: u-vector
$T \leftarrow$ Gabor filterbank (x)
Window length = 30 ms, window overlap = 15 ms
For j=1:number of frames **do**
$Var_t \leftarrow$ Variance (T(j, :), mean) $\leftarrow$**{t-vector}**
$mean_f \leftarrow$ mean(FFT$(T(j,:)), freq)/(2 * pi)$
$Var_f \leftarrow$ Variance(A, $mean_f$, freq) $\leftarrow$**{ω-vector}**
$tbp_{gen} \leftarrow var_t * var_f \leftarrow$**{$u$-vector}**
end for
return tbp_{gen}

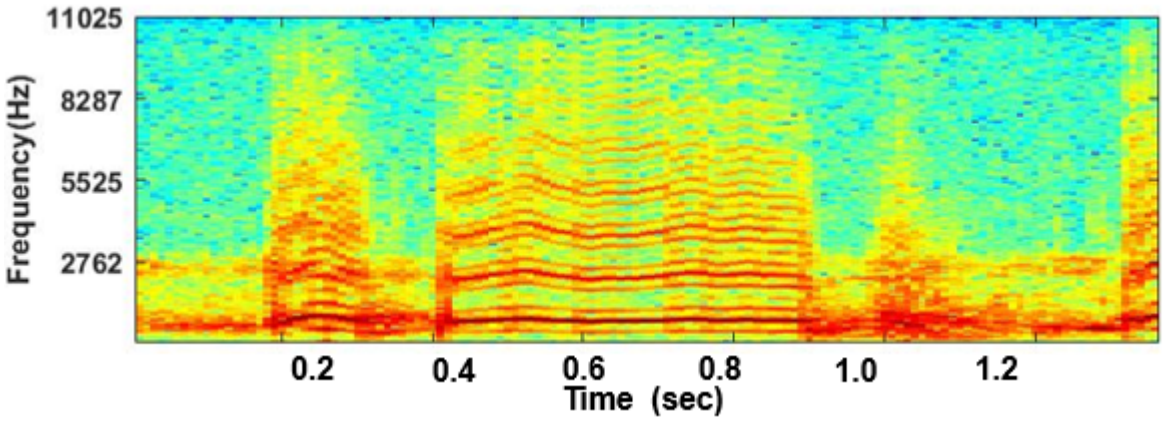
5.4 Experimental Results

We performed the experiments by fine-tuning feature parameter such as window overlap and number of subband filters. To that effect, we first varied the window overlap with values as 10, 15, and 20 ms. Number of subband filters were kept constant. The obtained experimental results are presented in Table 5.1.

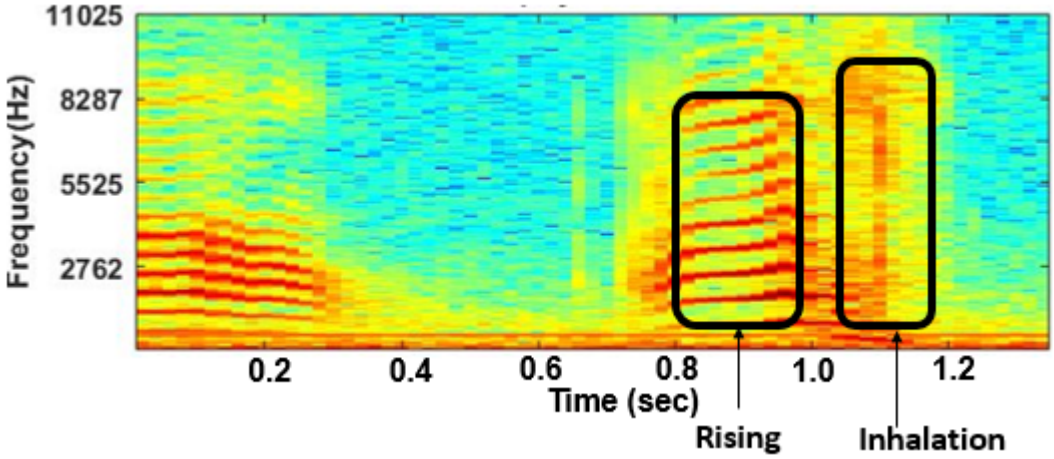
Table 5.1: % Accuracy for Spectral and Cepstral u-vector. After [7].

Window Length	Window Overlap	# Filters	% Accuracy (Spectral)	% Accuracy (cepstral)
30	15	40	**93.83**	93.04
30	15	60	93.08	**93.48**
30	15	80	87.71	87.00
30	20	40	93.35	92.42

From Table 5.1, it can be observed that the highest performance is achieved as 93.83% classification accuracy, obtained when window length, window overlap, and number of subband filters are of 30 ms, 15 ms and 40 ms respectively. Further, the next set of experiments were performed by varying the subband filters and keeping window overlap constant. These fine-tuning were performed considering the two cases of spectral, and cepstral u-vector. It should be noted that the spectral u-vector (with 93.83% classification accuracy) performs better than



(a)



(b)

Figure 5.2: Spectrograms of (a) Healthy *vs.* (b) Pathological Cries. After [7]

its cepstral version (with 93.48% classification accuracy). It can also be observed from Table 5.1 that as the number of subband filters increases, the % classification accuracy decreases.

Table 5.2: % Classification Accuracy for various Cepstral and spectral Feature Set. After [7].

Spectral Feature Set		**Cepstral Feature Set**	
Feature Set	**% Accuracy**	**Feature Set**	**% Accuracy**
u-vector	93.83	u-vector	93.48
t-vector	91.23	t-vector	89.38
ω-vector	**98.50**	ω-vector	96.74
CQT	97.00	CQT	**98.55**
Average	95.14	Average	94.53

Next, we compare the performance of u-vector, t-vector, and ω-vector with the

CQT baseline in Table 5.2. The comparison is done for both the cases of cepstral and spectral features. It can be observed that, the spectral ω-vector performs the best with % classification accuracy of 98.5% with overall increase of about 1.5% than the baseline CQT features. Hence, it can be observed that the frequency distribution patterns of the different cry modes, smeared over the entire frequency band, are captured by the ω-vector as discussed in [91].

Table 5.3: % Classification Accuracy of ω-vector with Various Number of Subband Filters. After [7].

Subband Filters	40	60	80	100
% Accuracy	**98.50**	91.37	92.20	96.78

Further, it can be observed that out of all the features shown in Table 5.2, the relatively best performance is achieved by ω-vector in the spectral case with an accuracy of 98.50%. Furthermore, it should also be noted that the average overall accuracy of spectral feature is higher than the cepstral features. In particular, the spectral features achieve average higher accuracy (95.14%) as compared to the cepstral features. This indicates that spectral features are better suited for pathology detection.

Given that ω-vector achieves the best performance in the spectral domain, we performed the next set of experiments to observe the effect of number of subband filters in the ω-vector. Table 5.3 presents the corresponding results, and it can be observed that the best result of 98.50% is achieved with 40 number of subband filters. From Table 5.3, we can say that when the entire frequency band is divided into 40 subbands, the frequency variance captured in each subband is optimum for our binary classification task.

5.5 Analysis of Latency Period

In this study, we also investigate the latency period for t-vector, ω-vector, and u-vector w.r.t the baseline CQT feature set considered in this study. The latency is estimated by the performance evaluation in terms of % accuracy w.r.t. varying duration of speech segment in an utterance. The duration of the utterance ranges from 20 ms to 600 ms, with an interval of 150 ms. Figure 5.3 shows comparison between spectral features of CQT, u-vector, t-vector, and ω-vector. It can be observed that the ω-vector outperforms u-vector and t-vector, and shows remarkable latency as compared to the CQT. Moreover, it can be observed that all the three features, i.e., u-vector, t-vector, and ω-vector gave increased % accuracy in a

short duration of speech utterance < 200 ms. On the other hand, CQT showed no improvement in accuracy even for a long duration of 600 ms of speech utterance. Further, the feature performance is better if for a low latency period the accuracy is high, which indicates the faster classification by the model and thus, indicating suitability for practical deployment of infant cry classification system.

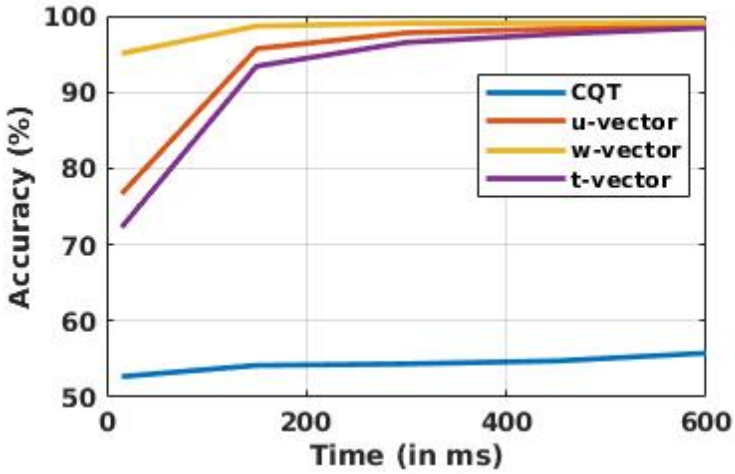


Figure 5.3: Latency Period *vs*. % Accuracy Between the Various spectral Features for CQT, u-vector, t-vector, and ω-vector. After [7].

5.6 Chapter Summary

In this study, u-vector (which is a combination of t-vector and ω-vector) is used to detect pathology from infant cries. These features are motivated from the Heisenberg's uncertainty principle in the signal processing framework for classifying between normal and pathological crying. To categorize the infant cries, this feature uses variances in time and frequency-domains, which correspond to t-vector and ω-vector. It is observed that ω-vector outperforms the remaining feature sets. This justifies the proposition that as compared to the healthy cries, pathological cries have an irregular frequency dispersion across the entire frequency band. Our experiments also show that spectral features are better suited for detection of pathological cries. Given the early detection of pathology in infants is also associated with faster detection, proposed features achieves relatively the best performance in latency.

CHAPTER 6
TECC For Speech Pathologies

6.1 Introduction

In this chapter, we discuss Teager Energy Cepstral Coefficients (TECC) implemented for the analysis and development of infant cry classification and dysarthric severity-level classification system. In Section 6.2, the brief description of the TEO and the extraction procedure is discussed. Further, in Section 6.3 the spectrographic analysis and TEO profile analysis are described through which the presence of linearities *vs.* non-linearities is investigated.

6.2 Proposed Feature Set

In the signal processing literature, energy of the speech signal $x(t)$ is estimated through L^2-norm of the signal, i.e., the integral of the square of absolute operation over the entire signal under analysis [71]. This method of estimating energy is based on linear filtering theory (in particular, Parseval's energy equivalence), which can describe *only* the linear components of speech production mechanism [67]. However, because the speech production mechanism is non-linear, the energy of the speech wave could not be effectively approximated using linear filter theory [86]. TEO was developed to address this problem [50]. It is a nonlinear differential operator that can capture the nonlinear feature of the speech production mechanism as well as the properties of the airflow pattern in the vocal tract system during speech production [71,78].

By approximating the derivative operation in continuous-time with backward difference in discrete-time, we obtain the TEO for a discrete-time signal $x(n)$ having amplitude, A and monocomponent angular frequency, Ω_m as follows [50]:

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1) \approx A^2\Omega_m^2. \tag{6.1}$$

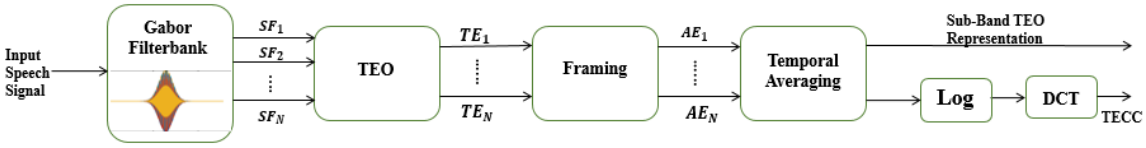


Figure 6.1: Functional Block Diagram of the Proposed Subband TEO representation and TECC Feature Set. (SF: Subband Filtered Signal, TE: Teager Energies, AE: Averaged Energies over frames). After [29,51].

TEO is derived to find the running estimate of the signal's energy for a monocomponent signal. However, speech signal consists of the frequency range varying from baseband to Nyquist frequencies. Hence, to obtain the monocomponent approximation of the signal, the speech signal is passed through the filterbank, which consists of several subband filters with appropriate center frequency and bandwidth. The subband filtered signals are narrowband signals, which are supposed to approximate the monotone signals and hence, TEO can be applied on these subband filtered signals. In this work, Gabor filterbank with linearly-spaced subband filters, is utilized for subband filtering. We chose Gabor subband filters due to their *optimal* time and frequency resolution in the framework of Heisenberg's uncertainty principle [67]. TEO is applied on each subband filtered signal to accurately estimate the energy. Furthermore, these narrowband energies are segmented into the frames of 20 *ms* duration with overlapping of 10 *ms*. Then, the temporal average for each frame is estimated to produce N-dimensional (D) *subband Teager energy representations (subband-TE)*. Discrete Cosine Transform (DCT) is performed on *subband Teager energy representations* to obtain the TECC feature set. The functional block diagram representation of the proposed subband-TE and TECC feature set is shown in Figure 7.1. Throughout this study, TECC features extracted using linear frequency scale are termed as TECC.

6.3 Experimental Analysis

6.3.1 Spectrographic Analysis of Infant Cry

In Fig. 6.2, Panel-I and Panel-II represents the spectrographic analysis for randomly sampled normal and asphyxia cry signals, respectively. Fig. 6.2(a), Fig. 6.2(b), and Fig. 6.2(c) represents the STFT, MelFB, and subband-TE representations, respectively. It can be observed from Fig. 6.2(a) that there is a difference in the pattern formed by F_0 and its harmonics for normal *vs.* asphyxia cry signals. These differences in the pattern are also visible for MelFB representation, as

shown in Fig. 6.2(b).

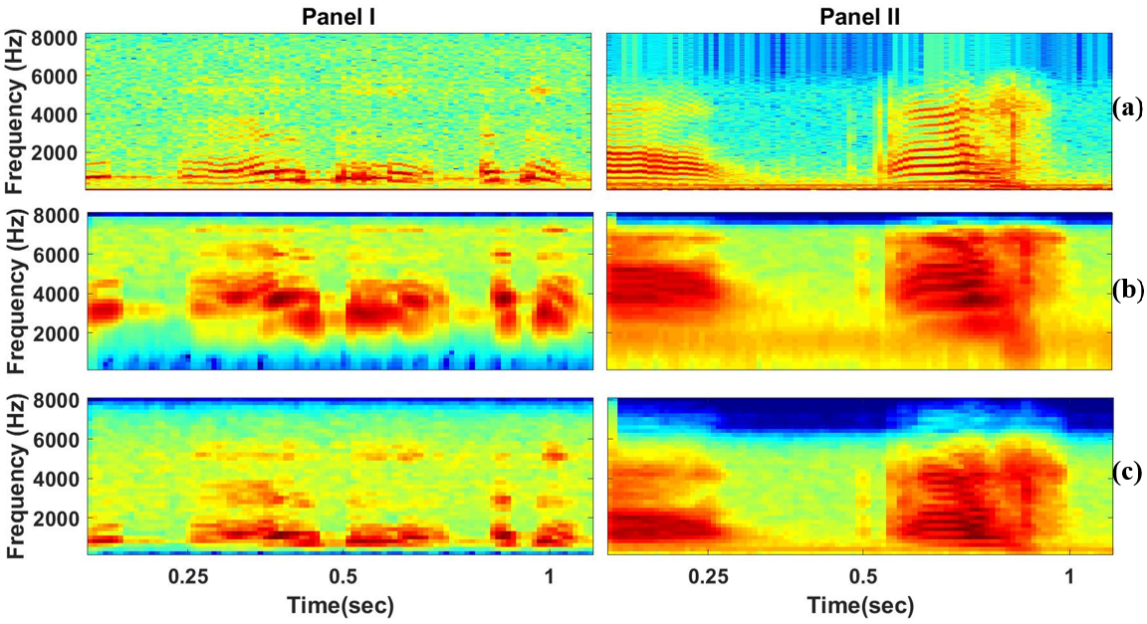


Figure 6.2: Panel-I and Panel-II represents the Spectrographic Analysis for Normal *vs.* Asphyxia Cry Samples, respectively. Fig. 6.2(a), Fig. 6.2(b), and Fig. 6.2(c) represents the STFT, MelFB, and Subband-TE Representations, respectively. After [9]

However, these differences are more vivid for subband-TE representations, as shown in Fig. 6.2(c). It might be because TEO can accurately estimate the energy of the signal, considering non-linear aspects of the speech production mechanism and also properties of airflow pattern in the vocal tract system [71, 78]. Furthermore, the results obtained using *10*-fold cross-validation also validates that the proposed TECC and subband-TE representations performs better over the other feature sets in this study.

6.3.2 Teager Energy Operator (TEO) Profile Analysis

Here, we analyse the TEO profiles around the 1^{st} formant frequency (i.e., $F1 = 500Hz$) for the utterance *w.r.t.* the same text material for normal *vs.* severity-levels. Panel-I of Figure 7.2 shows the subband filtered signal around 1^{st} formant frequency using a linear-spaced Gabor subband filter, and Panel-II shows corresponding TEO profiles.

Figure 7.2(a), Figure 7.2(b), Figure 7.2(c), Figure 7.2(d), and Figure 7.2(e) shows the analysis for normal, very low, low, medium, and high severity-levels, respectively. It can be observed that TEO profile for normal speech shows bumps within two consecutive Glottal Closure Instants (GCIs), which are known to in-

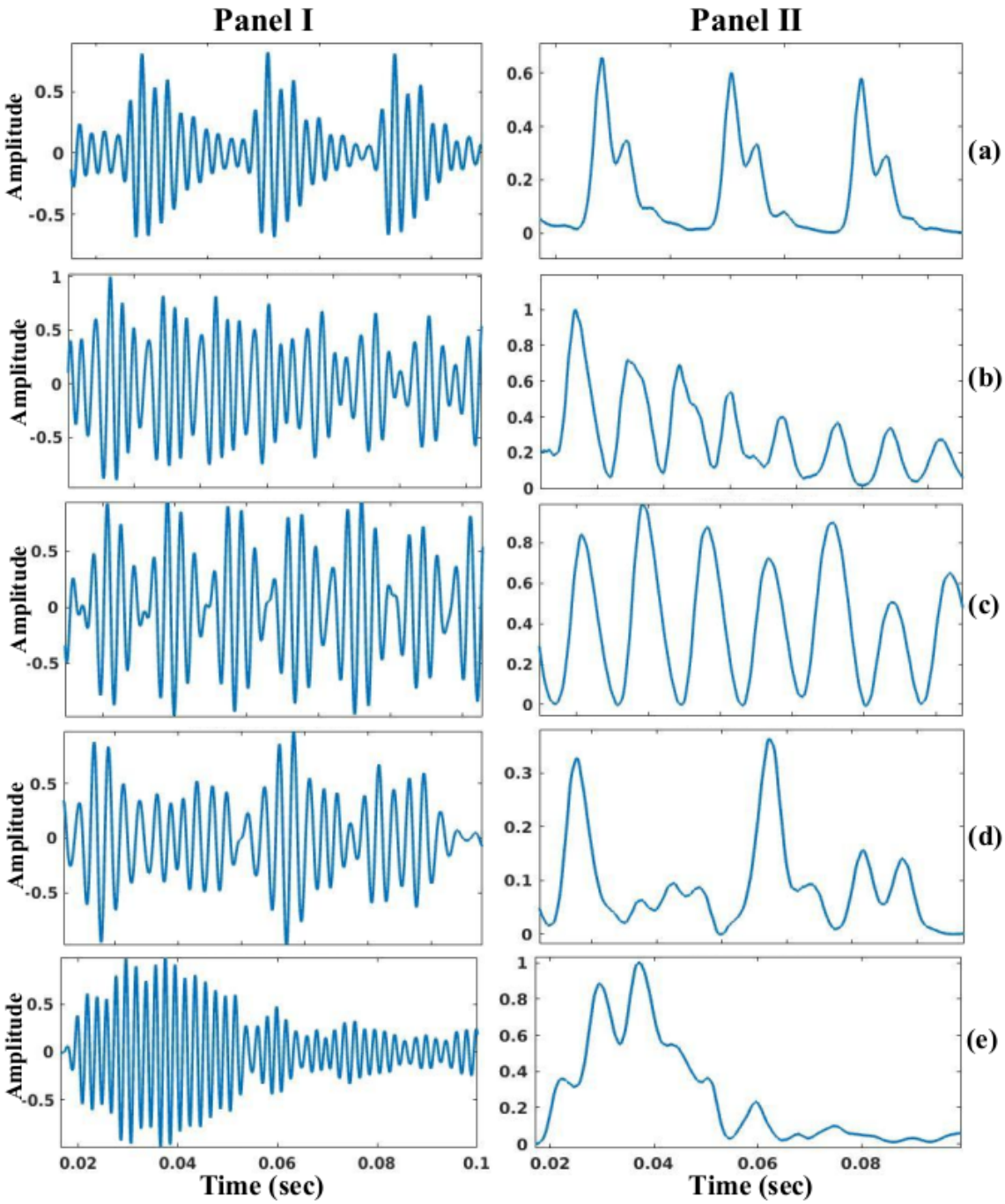


Figure 6.3: Subband Filtered Signal for Male Speakers Around $1^{st}Formant = 500Hz$ (Panel I) and corresponding TEO Profile (Panel II) for (a) Normal, Dysarthic Speech with Severity as (b) Very Low, (c) Low, (d) Medium, and (e) High. After [67].

dicate non-linearities in the speech production mechanism. Furthermore, it can also be observed that the quasi-periodicity in glottal excitation source decreases with increase in severity-level (as observed via aperiodic nature of TEO profile) indicating *disruption* in the rhythmic quasi-periodic movements of the vocal folds

due to dysarthria. Moreover, it is all the more significant in high dysarthric condition. Furthermore, as the severity-level increases, the neuro-motor impairment also increase, which leads to increased vocal fold closure disruption and loosing structural periodicity.

6.4 Experimental Results

Results on Baby Chilanto Database

Cepstral representations are being common in speech signal processing applications, we performed the experiments using four cepstral feature sets, namely, MFCC, LFCC, STCC, and TECC. As the size of the database is relatively small, experiments are performed using *10*-fold cross-validation. The database consists of the healthy and pathological class infant cry samples recorded with sampling rate of 22 kHz and 11 kHz, respectively. The experiments are performed using features extracted from the cry samples resampled to 16 kHz and results are reported in Table 6.1. It can be observed that the proposed TECC feature set outperforms the other feature sets for both SVM and GMM classifiers. We utilized 512 Gaussian mixtures in the GMMs. Furthermore, experiments are extended with spectral feature sets, namely, subband-TE, MelFB, LinFB, and STFT. We utilized the spectral feature representations as it has low-dimensional representations than the cepstral features. It can be observed from Table 6.2 that the proposed subband-TE feature set outperforms the other feature sets for both SVM and GMM classifiers. Furthermore, all the spectral representations performs equally well as compared to their corresponding cepstral representations. However, subband-TE performs slightly better than its cepsral counterpart, i.e., TECC. Hence, it would be better to choose the spectral representations for this application.

Table 6.1: Results for Various Cepstral Feature Sets. After [9].

		MFCC	LFCC	STCC	TECC
GMM	Acc.	98.55	98.28	98.99	**99.12**
	EER	1.23	0.50	0.26	0.61
SVM	Acc.	88.11	80.18	87.84	86.56
	EER	12.72	18.78	13.84	12.57

Table 6.2: Results for Various Spectral Feature Sets on GMM and SVM. After [9].

		MelFB	LinFB	STFT	subband-TE
GMM	Acc.	98.99	98.77	98.59	**99.47**
	EER	1.5	0.70	1.6	**0.3678**
SVM	Acc.	88.15	87.80	78.06	**90.35**
	EER	10.49	10.40	19.41	**8.23**

Table 6.3: Results w.r.t. Various GMM Mixtures. After [9].

Mixtures in GMM	64	128	256	512	1024
Accuracy	98.72	98.94	99.16	**99.47**	99.47

Table 6.4: Results for various Subband Filters. After [9].

Filters	40	60	80	100	120	140	160	180
Acc.	**99.47**	99.21	99.47	99.38	99.47	99.38	99.38	99.47

Results on UA-Speech Corpus

The results obtained in % classification accuracy using various features sets and classifiers are reported in Table 7.1. It can be observed from the Table 7.1 that the TECC performs relatively better than the baseline with classification accuracy of 97.18%, 94.63%, and 98.02% (i.e., absolute improvement of 1.98%, 1.41%, and 1.69%) for CNN, LCNN, and ResNet classifiers, respectively.

Table 6.5: Results for Various Classification Systems. After [6].

Feature Set $\downarrow$	**% Classification Accuracy**		
	CNN	**LCNN**	**ResNet**
MFCC	95.20	93.22	96.33
LFCC	96.32	94.07	97.17
TECC-Mel	92.37	85.87	93.09
TECC	**97.12**	**94.63**	**98.02**
MelFB	96.04	91.24	97.45
LinFB	94.91	89.26	97.17
Subband-TE	95.48	93.22	95.12

Furthermore, it is observed that optimum results of TECC are obtained for linear frequency scale. As mentioned in [52], the cepstral features perform better on noisy signal. In [94], the noise in dysarthric speech increases with increase in severity-levels. Hence, experiments were also performed on the spectral features *w.r.t* proposed and baseline features with all the three classifiers. It was observed that the cepstral features gave remarkably better % classification accuracy on all the classifiers. Hence, it can be inferred that more the severity-level, more is the

speech production noise.

6.5 Performance Evaluation

6.5.1 Infant Cry

DET curves are plotted for various spectral features as shown in Fig. 6.4. It can be observed that the proposed subband-TE representation performs better than all the other spectral representations for both the classifiers. The experiments are extended for varying number of Gaussian mixtures in GMM and results are obtained as shown in Table 6.3. It can be observed that the performance is improving as we increase the number of Gaussian mixtures in GMM from 64 to 512 and then it saturates, possibly due to the fact that a large number of 1024 mixtures is not required to model relatively lesser duration of infant cry samples. Hence, we utilized 512 Gaussian mixtures in GMM for the remaining experiments. Furthermore, performance is also validated w.r.t. number of subband filters in the Gabor filterbank to extract the subband-TE representations, and the results are reported in Table 6.4. It can be observed that the performance is almost constant w.r.t. number of subband filters in the filterbank and hence, we chose 40 number of subband filters in the filterbank as an optimal choice.

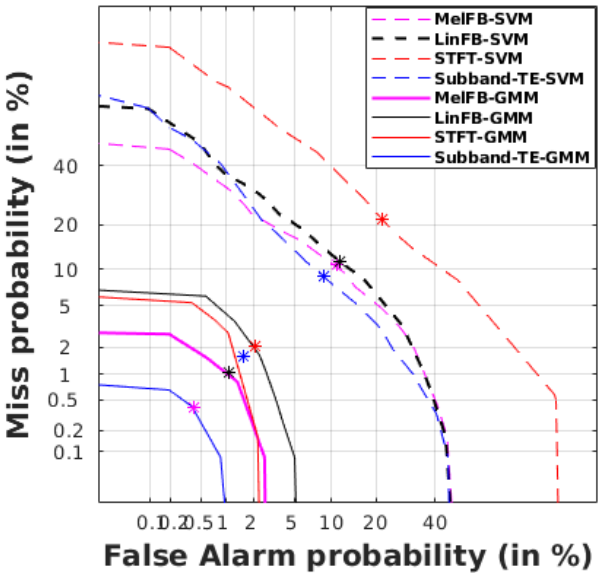


Figure 6.4: DET Plots for Different Feature Sets using Classifiers, namely, GMM and SVM for Infant Cry Classification. After [9].

6.5.2 Dysarthric Speech

Table 6.6 shows the confusion matrices for the TECC, MFCC, and LFCC for ResNet model. It can be observed that TECC reduces the misclassification errors, especially for high severity-level dysarthria, and overall performance of the TECC is relatively better than the MFCC, and LFCC. Furthermore, $F1$-score, MCC, Jaccard index, and Hamming loss are estimated for all the cepstral features as shown in Table 6.7. It can be observed from Table 6.7 that the TECC feature set outperforms the other cepstral features for all the evaluation metrics, indicating relatively better feature discriminative power of TECC.

Table 6.6: Confusion Matrix for MFCC, LFCC, and TECC using ResNet. After [6]

Feature	Severity	High	Medium	Low	Very Low
MFCC	**High**	72	0	2	1
	Medium	1	90	2	0
	Low	1	1	88	3
	Very Low	1	0	0	92
LFCC	**High**	74	0	1	0
	Medium	1	88	2	2
	Low	0	1	91	1
	Very Low	1	0	0	92
TECC	**High**	74	1	0	0
	Medium	1	92	0	0
	Low	0	1	92	0
	Very Low	1	0	0	92

Table 6.7: Various Statistical Measures for MFCC, LFCC, and TECC. After [6]

Feature Sets	$F1$-Score	MCC	Jaccard Index	Hamming Loss
MFCC	0.96	0.95	0.93	0.033
LFCC	0.97	0.96	0.95	0.025
TECC	**0.98**	**0.97**	**0.96**	**0.019**

Analysis of Latency Period

We analysed latency period for TECC, LFCC, and MFCC feature sets as shown in Figure 6.5. The latency period of the trained model is estimated by computing the % classification accuracy *w.r.t.* varying durations of test speech segment in

a test utterance. For latency period analysis, we chose the duration of the utterances varying from 100 ms to 3000 ms. The better performing model w.r.t. latency period should produce the larger accuracy for short speech segments. Moreover, it can be observed that the TECC gave significant % classification accuracy in a limited duration speech utterance of < 500 ms. On the contrary, MFCC and LFCC shows increment in accuracy after a relatively longer utterance duration of 1000 ms. Hence, these results signifies the suitability of TECC for practical dysarthric speech classification system deployment.

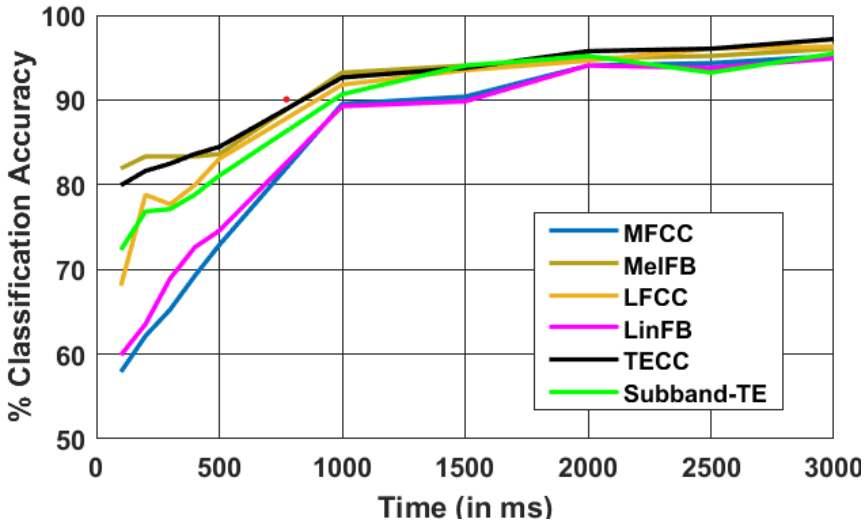


Figure 6.5: Latency period vs. % Classification Accuracy comparison between MFCC, MelFB, LFCC, LinFB, TECC, and Subband-TE. Best viewed in colour. After [6].

6.6 Chapter Summary

In this chapter, we investigated the use of Teager Energy-based features for classification of infant cry and dysarthric severity-level. It was observed that the spectral representation performs better for infant cry classification. This is due to the high pitch-source harmonics, spectral representations are more suitable for the normal *vs.* pathological infant cry classification. Whereas, it can be observed that the cepstral representation of the Teager energy performs best for dysarthric severity-level classification. This is due to the fact that the cepstral features perform better for noisy signals, and dysarthric speech is found to have production noise. This theoretical assumption is validated using the experimental results and analysis using spectrograms and DET curve. Moreover, it proposes the proposition that as the severity-level increases, the non-linearities decreases.

CHAPTER 7
Energy-Based Feature for Dysarthric Speech Analysis

7.1 Introduction

As discussed in the previous chapter, the energy-based features are capable of capuring better disctiminitive cues for dysarthric severity-level classification compared to auditory-based features. To validate this hypothesis, inthis chapter, we intoduce another-energy-based feature based on L^2 Norm of the signal. The Performance of L^2 Norm Energy Cepstral Coefficients (LECC) are compared with TECC and MFCC. State-of-the-art MFCC feature set is used as baseline for this study as in [48]. Formant enhancers are employed to enhance the dysarthric speech [49]. To that effect, this chapter presents the Formnat analysis for various dysarthric severity-level around the 1^{st} formant frequency for vowel $/e/$.

7.2 Proposed Work

In the signal processing literature, the energy of the speech signal $x(t)$ is estimated by calculating the integral of the square of absolute operation across the entire signal under consideration, i.e., estimating the L^2 norm of the signal, referred to as LEO [71]. This energy estimation method is based on linear filtering theory (specifically, Parseval's energy equivalence, the total energy of a signal, i.e., L^2 norm is conserved in the frequency-domain and this is also the condition of existence of inverse for several *linear* transforms, such as Fourier, Gabor, and Wavelet transforms), which can only represent the L^2 norm components of the speech generation process [67].

For LECC extraction, these narrowband output signals from Gabor filterbank are squared to estimate corresponding energies. Next, these narrowband energies are segmented with similar number of frames and window overlap. Temporal

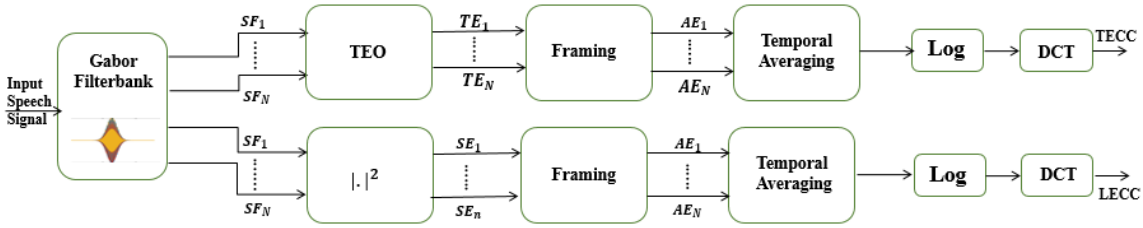


Figure 7.1: Functional Block Diagram of the Proposed TECC and LECC Feature Sets. (SF: Subband Filtered Signal, SE: Squared Linear Energies, TE: Teager Energies, AE: Averaged Energies over frames). After [29].

averaging for each frame is estimated (i.e., L^2 norm of each subband signal) to get N-D *subband L^2 norm Energy representation (subband-LE)*. Discrete Cosine Transform (DCT) is applied on *subband L^2 norm energy representations* in order to obtain the LECC. The functional block diagram representation of TECC and LECC feature sets is shown in Fig 7.1. Throughout this chapter, TECC and LECC features extracted using linear frequency scale and for both the feature sets, DCT does the job of feature decorrelation, energy compaction, and feature vector dimensionality reduction.

7.3 Analysis of LEO Profiles

Here, we analyse the TEO profiles around the 1^{st} formant frequency (i.e., $F_1 = 500Hz$) for the utterance *w.r.t.* the same text material for normal *vs.* severity-levels. Panel I of Fig. 7.2 shows the subband filtered signal around 1^{st} formant (F_1) frequency using a linearly-spaced Gabor subband filter, and Panel II shows corresponding TEO profiles. Fig. 7.2(a), Fig. 7.2(b), Fig. 7.2(c), Fig. 7.2(d), and Fig. 7.2(e) shows the analysis for normal, very low, low, medium, and high severity-levels, respectively. It can be observed that TEO profile for normal speech shows *bumps* within two consecutive Glottal Closure Instants (GCIs), which are known to indicate non-linearities in speech production mechanism [78]. Furthermore, it can also be observed that the quasi-periodicity in glottal excitation source decreases with increase in severity-level (as observed via aperiodic TEO profile) indicating *disruption* in the rhythmic quasi-periodic movements of the vocal folds due to dysarthria. Moreover, it is all the more significant in high severity-level dysarthric condition. Furthermore, as the severity-level increases, the neuro-motor impairment also increases, which leads to increased disruption in vocal fold closure and loosing *structural* periodicity. From Panel III of Fig. 7.2, which shows the LEO profiles around 1^{st} formant frequency for vowel /*e*/, it can be observed

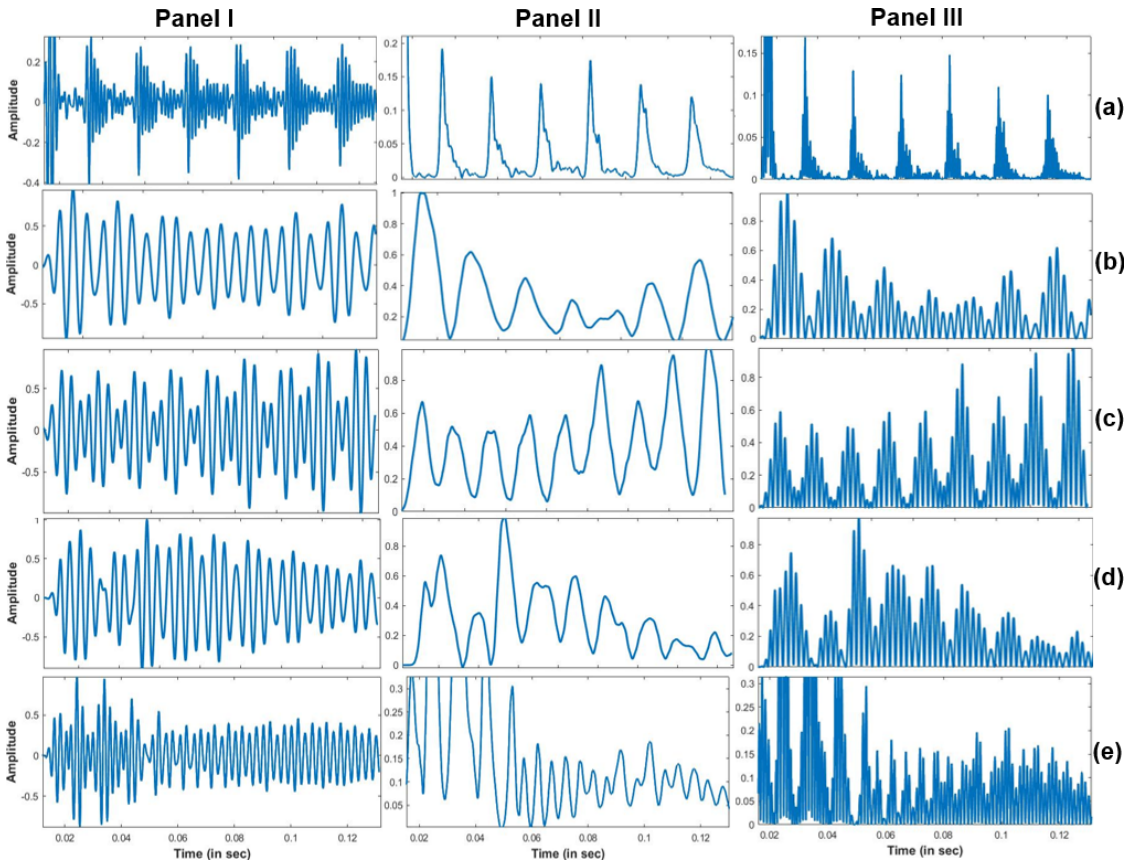


Figure 7.2: Subband Filtered Signal (for Vowel /e/) for Male Speakers Around 1^{st}Formant$F_1 = 500Hz$ (Panel I), Corresponding TEO Profile (Panel II), and Corresponding $|.|^2$ Envelope (Panel III) for (a) Normal, Dysarthic Speech with Severity-Level as (b) Very Low, (c) Low, (d) Medium, and (e) High. After [67].

that the LEO is capable of maintaining the periodicity in the speech produced by dysarthric speaker, which are not captured by TEO due to possible decrease in non-linearities. Hence, it can be said that as the dysarthric severity-level increases, the linearities in speech signal increases.

7.4 Experimental Results

The results obtained as % classification accuracy using various feature sets are reported in Table 7.1. It can be observed that LECC performs relatively better than the baseline MFCC with classification accuracy of 1.7% (4.23%) on CNN (LCNN) classifier systems, respectively. Furthermore, LECC performs better than the baseline MFCC explored in [48]. The analysis in the subsequent section, along with the classification accuracy obtained using various classifiers, indicate that the energy-based features are capabale of capturing better discriminative cues for dysarthric severity-level classification rather than auditory-based features.

Table 7.1: Results For various Classification Systems. After [5].

Feature Set	% Classification Accuracy	
	CNN	LCNN
MFCC	96.32	92.09
LECC	**98.02**	**96.32**

7.4.1 Performance Evaluation

The Table 7.2 shows the confusion matrices for MFCC and LECC. It can be observed that LECC reduces the misclassification errors corresponding to the different severity-levels, indicating the better performance of LECC *w.r.t.* MFCC. Furthermore, performance of LECC *w.r.t.* MFCC is also analysed using $F1$-Score, MCC, Jaccard Index, and Hamming Loss as shown in Table 7.3. It can be observed from Table 7.3 that LECC performs better than the MFCC for the dysarthic severity-level classification.

Table 7.2: Confusion Matrix for MFCC and LECC using CNN. After [5].

Feature	Severity	High	Medium	Low	Very Low
MFCC	**High**	67	4	3	1
	Medium	2	90	0	0
	Low	1	1	91	0
	Very Low	1	0	0	92
LECC	**High**	74	1	0	0
	Medium	2	90	0	0
	Low	1	0	92	0
	Very Low	0	0	0	93

Table 7.3: Various Statistical Measures of MFCC and LECC. After [5].

Feature Set	F1-Score	MCC	Jaccard Index	Hamming Loss
MFCC	0.96	0.95	0.82	0.036
LECC	**0.98**	**0.97**	**0.96**	**0.019**

7.4.2 Linear Discriminate Analysis (LDA)

The capability of LECC to classify severity-level is also validated by LDA scatter plots, which projects the higher-dimensional feature space to the lower-dimension [46]. Here MFCC and LECC features are projected to the 2-D space to get the

scatter plots for various severity-levels of dysarthria. Fig. 7.3(a) and Fig. 7.3(c) shows the LDA plots of MFCC and LECC, respectively. From the Fig. 7.3, it can be observed that for LECC, the variance of each severity-level clusters is less resulting in relatively better performance of LECC, which increases the interclass distance between the clusters than the :MFCC and TECC.

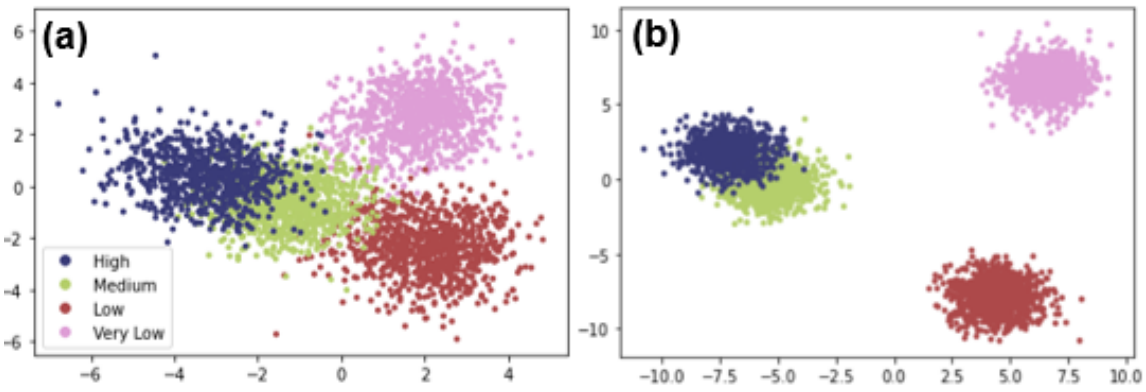


Figure 7.3: Scatter Plots obtained using LDA for (a) MFCC and (b) LECC. After [46]. Best viewed in colour.

7.5 Chapter Summary

The effectiveness of the energy-based features *vs.* auditory-based features on the analysis and classification of the severity-level of dysarthric speech was analysed in this chapter. To validate the effect of energy in dysarthric speech and severity-level classification, the L^2 norm energy operator, i.e., L^2 norm is analysed against TEO. The TEO profile of a normal speech signal shows bumps that represent non-linearities in the speech production mechanism. This bumpy structure, on the other hand, diminishes as the severity-level increases. Experiments with CNN and LCNN classifiers are used to test these hypotheses. The experimental results demonstrated that the L^2 norm operator is more suitable for dysarthric speech analysis and classification than auditory based features like MFCC. Statistical measures, such as the $F1$-score, MCC, Jaccard Index, Hamming Loss, and LDA were used to validate the observation.

CHAPTER 8
Summary and Conclusions

8.1 Summary

The work presented in this thesis aims towards development of effective methods for classification of infant cry and dysarthric severity-level. The practical potential of various signal processing based approaches are studied in this work. Apart from the signal processing approaches, various deep-learning classifiers, and it's effect on dysarthric severity-level classification is also studied in this chapter. In addition to this, a comparative study is also done for spectral *vs.* cepstral features for the classification of pathological cries and dysarthric severity-level classification. A comparative study between energy-based features such as TECC, LECC and perception-based features such as MFCC, and LFCC for capturing the effective discriminative cues in infant cry and dysarthric speech is also studied in this work. In addition to this, the ten cry modes in infant cries are also studied. Further, the form-invariance property of CQT was studied in this work. The effectiveness of CQT and it's form-invariance property was tested under the signal degradation condition for babble and car noise. The motor control disorder in dysarthric patients against a normal person was analysed around the 1^{st} formant frequency of vowel $/i/$ using TEO and following its analysis, the same analysis was done using LEO around the 1^{st} formant frequency of vowel $/e/$. Finally, the performance of each model was evaluated using various evaluation metrics.

8.2 Limitations of the Current Work

Although our model for classification of infant cry analysis and dysarthric severity-level classification gave remarkable results, following are some limitations of our work as follows:

- The availability of different pathological cries is always an ordeal when it

comes to quantitative perspective.

- Although certain pathological cries are available in various databases, however all the pathologies are not available in all databases. This creates a challenge for cross database evaluation.
- Presence of linear component in dysarthric speech is still not generalized, as very limited work has been explored in terms of signal processing framework.
- Cross database evaluation has not been explored in terms of either signal processing framework or in DNN perspective due to lack of standardization when it comes to severity-levels in different database.
- The motor control distortion between brain and primary speech mechanism or brain and secondary speech producing mechanism remains an unexplored region of research.

8.3 Future Research Directions

Based on the limitations of our work, we present the future work directions:

- To overcome the challenge of limited dataset, various data augmentation techniques can be applied on available database for infant cry analysis and classification. Apart from this, the effect of various parameters like pitch, speed, and tempo can also be analysed in infant cries.
- Cross database evaluation for infant cry analysis can be conducted for certain class of pathologies which are common across available databases using traditional machine learning as well as modern deep learning methods.
- Autism spectrum disorder classification in the infant cries is a prominent research direction with only hindrance being limited database.
- Multi-class classification of pathological cries based on type of pathologies.
- Analysis of presence of linearities in dysarthric speech using various signal processing frameworks on cross database evaluation.
- Source glottal filtering method can be used to analyse the formant frequency and know the most effect source of speech production for dysarthric patients.

- An ensemble of signal processing and deep learning based approach can be implemented for dysarthric speech enhancement, that can be used for various applications such as voice assistants and ASR systems.
- A system can be designed to recognize the unintelligible words produced by dysarthric patients inherently. Hence developing the ASR models for dysarthric patients.

CHAPTER 9
List of Publications

(1) Hemant A. Patil, Ankur T. Patil, **Aastha Kachhi**, "Constant Q Cepstral Coefficients for Classification of Normal *vs.* Pathological Infant Cry," **published in**: International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22-27 May 2022, pages 5.

(2) Ankur T. Patil, **Aastha Kachhi**, Hemant A. Patil, "Subband Teager Energy Representations for Infant Cry Analysis and Classification", **accepted in**: European Signal Processing Conference (EUSIPCO), Belgrade, 29 Aug-02 Sept 2022.

(3) **Aastha Kachhi**, Priyanka Gupta, Hemant A. Patil, "Features Motivated From Uncertainty Principle for Classification of Normal *vs.* Pathological Infant Cry", **accepted in**: European Signal Processing Conference, (EUSIPCO), Belgrade, 29 Aug-02 Sept 2022.

(4) Anand Therattil, **Aastha Kachhi**, Hemant A. Patil, "Cross-Teager Energy Ceptsral Coefficients For Dysarthric Severity-Level Classification", **accepted in**: INTERSPEECH Workshop, Korea, 18-22 Sept 2022.

(5) Hemant A. Patil, Ankur T. Patil, **Aastha Kachhi**, "On Significance of Constant-Q Transform for Infant Cry Analysis and Classification," **rejected in:** INTERSPEECH, Korea, 18-22 Sept 2022.

(6) **Aastha Kachhi**, Anand Therattil, Ankur T. Patil, Hardik B. Sailor, Hemant A. Patil, "Dysarthric Speech Severity-Level Analysis and Classification Using Teager Energy Cepstral Features", **rejected in:** INTERSPEECH, Korea, 18-22 Sept 2022.

(7) **Aastha Kachhi**, Anand Therattil, Ankur T. Patil, Hardik B. Sailor, Hemant A. Patil, "Analysis of Non-Linearities in Normal *vs.* Dysarthric Speech For

Severity-Level Classification", **rejected in:** International conference on Signal Processing and Communications (SPCOM), IISc Bangalore, 11-15 July 2022.

(8) **Aastha Kachhi**, Anand Therattil, Priyanka Gupta, Hemant A. Patil, "Continuous Wavelet Transform for Severity-Level Classification of Dysarthria", **rejected in:** International conference on Signal Processing and Communications (SPCOM), IISc Bangalore, 11-15 July 2022.

(9) Hemant A. Patil, Ankur T. Patil, **Aastha Kachhi**, Anand Therattil "Novel Constant-Q Cepstral Features for Infant Cry Classification," **rejected in:** International conference on Signal Processing and Communications (SPCOM), IISc Bangalore, 11-15 July 2022.

References

[1] Baby Crying Analyzer, url: :http://www.showeryourbaby.com/whycrbacran1.html, note = Accessed: 2022-03-29.

[2] Baby Pod, howpublished = https://babypod.net/en/babypod-device/, note = Accessed: 2022-03-29.

[3] Medical Records, howpublished = ://www.medicalrecords.com/health-a-to-z/respiratory-system-in-a-child-multimedia/. Accessed: 2022-05-29.

[4] H. A. Patil, A. T. Patil, and A. Kachhi. Constant Q Cepstral Coefficients for normal *vs.* pathological infant cry. In *accepted in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022.

[5] A. P. H. B. S. H. A. P. Aastha Kachhi, Anand Therattil. Analysis of non-linearities in normal *vs.* dysarthric speech for severity-level classification. *SPCOM, IISC, Banglore, India*, 2022.

[6] A. P. H. B. S. H. A. P. Aastha Kachhi, Anand Therattil. Dysarthric speech severity-level analysis and classification using teager energy cepstral features. *Interspeech*, 2022.

[7] H. A. P. Aastha Kachhi, Priyanka Gupta. Features motivated from uncertainty principle for classification of normal *vs.* pathological infant cry. *European Signal Processing Conference*, 2022.

[8] H. F. Alaie, L. Abou-Abbas, and C. Tadj. Cry-based infant pathology classification using gmms. *Speech Communication*, 77:28–52, 2016.

[9] H. A. P. Ankur Patil, Aastha Kachhi. Subband teager energy representations for infant cry analysis and classification. *European Signal Processing Conference*, 2022.

[10] V. Apgar. A proposal for a new method of evaluation of the newborn. *Classic Papers in Critical Care*, 32(449):97, 1952.

[11] L. Armbrüster, W. Mende, G. Gelbrich, P. Wermke, R. Götz, and K. Wermke. Musical intervals in infants' spontaneous crying over the first 4 months of life. *Folia Phoniatrica et Logopaedica*, 73(5):401–412, 2021.

[12] M. D. Beecher. Spectrographic analysis of animal vocalizations: Implications of the "uncertainty principle. *Bioacoustics*, 1(2-3):187–208, 1988.

[13] I.-A. Bğnicğ, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu. Baby cry recognition in real-world conditions. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 315–318. IEEE, 2016.

[14] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] B. Boashash. Estimating and interpreting the instantaneous frequency of a signal. *I.* fundamentals. *Proceedings of the IEEE*, 80(4):520–538, 1992.

[16] B. Boashash. Time-frequency and instantaneous frequency concepts (chapter 1). In *Time-Frequency Signal Analysis and Processing (Second Edition)*, pages 31 – 63. Academic Press, Oxford, 2016.

[17] Boashash, Boualem. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. Academic Press, Second Edition, 2015.

[18] M. Bouchard, A.-L. Jousselme, and P.-E. Doré. A proof for the positive definiteness of the Jaccard index matrix. *International Journal of Approximate Reasoning*, 54(5):615–626, 2013.

[19] J. C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America (JASA)*, 89(1):425–434, 1991.

[20] N. Buddha and H. A. Patil. Corpora for analysis of infant cry. 2007.

[21] P. Busch, T. Heinonen, and P. Lahti. Heisenberg's uncertainty principle. *Physics Reports*, 452(6):155–176, 2007.

[22] C.-Y. Chang and L.-Y. Tsai. A cnn-based method for infant cry detection and recognition. In *Workshops of the International Conference on Advanced Information Networking and Applications*, pages 786–792. Springer, 2019.

[23] Z. Chen, Z. Xie, W. Zhang, and X. Xu. Resnet and model fusion for automatic spoofing detection. In *INTERSPEECH*, pages 102–106, Stockholm, Sweden, August 2017.

[24] A. Chittora and H. A. Patil. Classification of pathological infant cries using modulation spectrogram features. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 541–545. IEEE, 2014.

[25] A. Chittora and H. A. Patil. Data collection of infant cries for research and analysis. *Journal of Voice*, 31(2):252–e15, 2017.

[26] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, (3):326–334, 1965.

[27] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. Regret analysis for performance metrics in multi-label classification: the case of Hamming and subset zero-one loss. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 280–295. Springer, 2010.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[29] D. Dimitriadis, P. Maragos, and A. Potamianos. Auditory Teager energy cepstrum coefficients for robust speech recognition. In *INTERSPEECH*, pages 3013–3016, Lisbon, Portugal, Sept. 2005.

[30] G. R. Doddington. Speaker recognition—identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985.

[31] M. Dorsey, K. Yorkston, D. Beukelman, and M. Hakel. Speech intelligibility test for windows. *Institute for Rehabilitation Science and Engineering at Madonna*, 2007.

[32] P. Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980.

[33] J. J. Engelsma, D. Deb, K. Cao, A. Bhatnagar, P. S. Sudhish, and A. K. Jain. Infant-ID: Fingerprints for global good. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[34] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[35] G. Z. Felipe, R. L. Aguiar, Y. M. Costa, C. N. Silla, S. Brahnam, L. Nanni, and S. McMurtrey. Identification of infants' cry motivation using spectrograms.

In *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 181–186. IEEE, 2019.

[36] J. L. Flanagan. *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media, 2013.

[37] D. Gabor. Theory of communication-part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[38] G. Gambardella. Time scaling and short-time spectral analysis. *The Journal of the Acoustical Society of America (JASA)*, 44(6):1745–1747, 1968.

[39] G. Gambardella. A contribution to the theory of short-time spectral analysis with nonuniform bandwidth filters. *IEEE Transactions on Circuit Theory*, 18(4):455–460, 1971.

[40] G. Gambardella. The Mellin transforms and constant-q spectral analysis. *The Journal of the Acoustical Society of America (JASA)*, 66(3):913–915, 1979.

[41] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[42] R. González Hautamäki, V. Hautamäki, and T. Kinnunen. On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. *The Journal of the Acoustical Society of America (JASA)*, 146(1):693–704, 2019.

[43] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117, 2021.

[44] A. K. Hemant A. Patil, Ankur Patil. On significance of constant-q transform for infant cry analysis and classification. *Interspeech*, 2022.

[45] A. T. P. P. G. Hemant A. Patil, Rajul Acharya. Non-cepstral uncertainty vector for replay spoofed speech detection. *European Signal Processing Conference*, 2022.

[46] A. J. Izenman. Linear discriminant analysis. In *Modern Multivariate Statistical Techniques*, pages 237–280. Springer, 2013.

[47] C. Ji, T. B. Mudiyanselage, Y. Gao, and Y. Pan. A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–17, 2021.

[48] A. A. Joshy and R. Rajan. Automated dysarthria severity classification using deep learning frameworks. In 28^{th} *European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands*, pages 116–120, 2021.

[49] A. B. Kain, J.-P. Hosom, X. Niu, J. P. Van Santen, M. Fried-Oken, and J. Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, 2007.

[50] J. F. Kaiser. On a simple algorithm to calculate the energy of a signal. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 381–384, New Mexico, USA, 1990.

[51] M. R. Kamble and H. A. Patil. Detection of replay spoof speech using teager energy feature cues. *Computer Speech & Language*, 65:101140, 2021.

[52] G. Korvel, O. Kurasova, and B. Kostek. Comparative analysis of spectral and cepstral feature extraction techniques for phoneme modelling. In *International Conference on Multimedia and Network Information System, Wroclaw, Poland*, pages 480–489. Springer, 2018.

[53] L. Cohen. *Time-Frequency Analysis*, volume 778. 1^{st} Edition, Prentice-Hall, 1995.

[54] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE Int. Symp. on Circuits and Systems*, pages 253–256, Paris, France, 2010.

[55] B. M. Lester and C. Z. Boukydis. No language but a cry. *Nonverbal vocal communication. Comparative and developmental approaches*, pages 145–73, 1992.

[56] P. Lieberman. Primate vocalizations and human linguistic ability. *The Journal of the Acoustical Society of America (JASA)*, 44(6):1574–1584, 1968.

[57] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients. *Journal of Speech and hearing Disorders*, 43(1):47–57, 1978.

[58] H. C. Mahendru. Quick review of human speech production mechanism. *International Journal of Engineering Research and Development*, 9(10):48–54, 2014.

[59] S. G. Mallat. Time Meets Frequency. In *A Wavelet Tour of Signal Processing*. Academic Press, Boston, 3^{rd} edition, 2009.

[60] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, 2^{nd} Edition, 1999.

[61] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *EUROSPEECH*, pages 1895–1898, Rhodes, Greece, 1997.

[62] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[63] J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amiel-Tison. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178, 1988.

[64] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[65] A. Messaoud and C. Tadj. A cry-based babies identification system. In *International Conference on Image and Signal Processing*, pages 192–199. Springer, 2010.

[66] L. Meteyard, S. R. Cuadrado, B. Bahrami, and G. Vigliocco. Coming of age.

[67] H. M.Teager and S. M.Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *William J. Hardcastle and Alain Marchal (Eds.), Speech Production and Speech Modelling*, pages 241–261. Springer, 1990.

[68] M. Nicolao, H. Christensen, S. Cunningham, P. Green, and T. Hain. A framework for collecting realistic recordings of dysarthric speech-the homeservice corpus. In *Proceedings of LREC 2016*. European Language Resources Association, 2016.

[69] C. C. Onu, I. Udeogu, E. Ndiomu, U. Kengni, D. Precup, G. M. Sant'Anna, E. Alikor, and P. Opara. Ubenwa: Cry-based diagnosis of birth asphyxia. *arXiv preprint arXiv:1711.06405*, 2017.

[70] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals and Systems*. Prentice Hall, 2^{nd} Edition, 1997.

[71] A. V. Oppenheim, A. S. Willsky, S. H. Nawab, G. M. Hernández, et al. *Signals & Systems*. Pearson Educación, 1997.

[72] R. Pandita. Spectrographic analysis of dysarthric speech. *Journal of All India Institute of Speech and Hearing*, 14(1):135–142, 1983.

[73] J. J. Parga, S. Lewin, J. Lewis, D. Montoya-Williams, A. Alwan, B. Shaul, C. Han, S. Y. Bookheimer, S. Eyer, M. Dapretto, et al. Defining and distinguishing infant behavioral states using acoustic cry analysis: is colic painful? *Pediatric research*, 87(3):576–580, 2020.

[74] H. A. Patil. Infant identification from their cry. In *2009 Seventh International Conference on Advances in Pattern Recognition*, pages 107–110, 2009.

[75] H. A. Patil. "cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry. In *Advances in speech recognition*, pages 323–348. Springer, 2010.

[76] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens. A new, robust vocal fundamental frequency (f/sub 0/) determination method for the analysis of infant cries. In *Proceedings of IEEE Symposium on Computer-Based Medical Systems (CBMS)*, pages 223–228. IEEE, 1994.

[77] M. Petroni, M. Malowany, C. Johnston, and B. Stevens. A crosscorrelation-based method for improved visualization of infant cry vocalizations. *Electrical and Computer Engineering*, 1994.

[78] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, 3^{rd} edition, India, 2006.

[79] A. F. Ribeiro and K. Z. Ortiz. Populational profile of dysarthric patients assisted in a tertiary hospital. *Revista da Sociedade Brasileira de Fonoaudiologia*, 14(4):446–453, 2009.

[80] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi. Classifying infant cry patterns by the genetic selection of a fuzzy model. *Biomedical Signal Processing and Control*, 17:38–46, 2015.

[81] F. Rudzicz, A. K. Namasivayam, and T. Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541, 2012.

[82] H. B. Sailor, M. R. Kamble, and H. A. Patil. Auditory filterbank learning for temporal modulation features in replay spoof speech detection. In *INTERSPEECH*, pages 666–670, Hyderabad, India, Sept. 2018.

[83] M. R. Schädler and B. Kollmeier. Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. *The Journal of the Acoustical Society of America (JASA)*, 137(4):2047–2059, 2015.

[84] M. R. Schroeder and B. S. Atal. Generalized short-time power spectra and autocorrelation functions. *The Journal of the Acoustical Society of America (JASA)*, 34(11):1679–1683, 1962.

[85] K. Sharma, C. Gupta, and S. Gupta. Infant weeping calls decoder using statistical feature extraction and gaussian mixture models. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2019.

[86] H. M. Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):599–601, 1980.

[87] R. I. Tuduce, M. S. Rusu, C. Horia, and C. Burileanu. Automated baby cry classification on a hospital-acquired baby cry database. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pages 343–346. IEEE, 2019.

[88] O. Wasz-Höckert, T. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne. The identification of some specific meanings in infant vocalization. *Experientia*, 20(3):154–154, 1964.

[89] K. Wu, C. Zhang, X. Wu, D. Wu, and X. Niu. Research on acoustic feature extraction of crying for early screening of children with autism. In *2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 290–295. IEEE, 2019.

[90] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[91] Q. Xie, R. K. Ward, and C. A. Laszlo. Determining normal infants' level-of-distress from cry sounds. In *Proceedings of Canadian Conference on Electrical and Computer Engineering*, pages 1094–1096. IEEE, 1993.

[92] Q. Xie, R. K. Ward, and C. A. Laszlo. Automatic assessment of infants' levels-of-distress from the cry signals. *IEEE Transactions on Speech and Audio Processing*, 4(4):253, 1996.

[93] K. M. Yorkston, D. R. Beukelman, and C. Traynor. *Assessment of intelligibility of dysarthric speech*. Pro-ed Austin, TX, 1984.

[94] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, K. H. Wong, X. Liu, and H. Meng. Development of the CUHK dysarthric speech recognition system for the UA speech corpus. In *INTERSPEECH, Hyderabad, India*, pages 2938–2942, 2018.

[95] A. Zabidi, W. Mansor, L. Y. Khuan, R. Sahak, and F. Rahman. Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In *2009 5th International Colloquium on Signal Processing & Its Applications*, pages 204–208. IEEE, 2009.

[96] A. Zabidi, W. Mansor, L. Y. Khuan, I. M. Yassin, and R. Sahak. Classification of infant cries with hypothyroidism using multilayer perceptron neural network. In *2009 IEEE International Conference on Signal and Image Processing Applications*, pages 246–251. IEEE, 2009.