Phase Based Methods for Various Speech Applications

by

Aditya Pusuluri
202115008

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

MASTER OF TECHNOLOGY
in
ELECTRONICS AND COMMUNICATION

with specialization in
Communication Systems and Machine Learning
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, 2023

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.

P.S.Aditya

Aditya Pusuluri

Certificate

This is to certify that the thesis work entitled Classification of Dysarthric Speech has been carried out by Aditya Pusuluri 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 my/our supervision.

H Patil

Prof. Hemant A. Patil
Thesis Supervisor

Acknowledgments

As I reach the culmination of my thesis, I am filled with a profound sense of gratitude for the unwavering support and encouragement I have received from numerous individuals who have played a pivotal role in my academic journey. The completion of this work would not have been possible without their steadfast belief in me, especially during the unprecedented challenges posed by the pandemic. First and foremost, I extend my deepest appreciation to my supervisor, Prof. Hemant A. Patil. His expertise, guidance, and unwavering support have been instrumental in shaping the research direction and ensuring its successful completion. His constructive feedback, patience, and dedication to my academic growth have been a constant source of inspiration and motivation. I am immensely grateful to the Speech Research Lab at DAIICT for providing the necessary resources, infrastructure, and collaborative environment that have fostered my research endeavours. The contributions of the lab members, both past and present, have played a significant role in shaping my understanding and passion for speech research. I extend my gratitude to each member for their support and valuable insights.

I would like to express a special thanks to Ms. Aastha Kachhi, an alumnus of the Speech Research Lab. Her mentorship, guidance, and willingness to share her expertise have been invaluable in enhancing my research skills and broadening my understanding of the field. Her encouragement and support have been pivotal in overcoming the challenges faced during the pandemic. I would also like to acknowledge the support and understanding of my family and friends Ms. Uthiraa, Mr. Prabhanshu Yadav, Mr. Avinash Subramaniam, and Mr. Arunangshu Dutta throughout this journey. Their unwavering belief in my abilities and their constant encouragement has provided me with strength and determination.

Finally, I extend my thanks to the PRISM team at Samsung Research Institute, Bangalore (SRI-B) and Project BASHINI, funded by the Ministry of Electronics and Information Technology(Meity) for providing me with the opportunity to build and finish my thesis.

Contents

Abstract

Vocal communication plays a fundamental role in human interaction and expression. Right from the first cry to adult speech, the signal conveys information about the well-being of the individual. Lack of coordination between the speech muscles and the brain leads to voice pathologies. Some pathologies related to infants are Asphyxia, Sudden Death Syndrome (SIDS), etc. The other voice pathologies that affect the speech production systems are dysarthria, cerebral palsy, and parkinson's disease.

Dysarthria, a neurological motor speech disorder, is characterized by impaired speech intelligibility that can vary across severity-levels. This works focuses on exploring the importance of Modified Group Delay Cepstral Coefficients (MDGCC)-based features in capturing the distinctive acoustic characteristics associated with dysarthric severity-level classification, particularly for irregularities in speech. Convolutional Neural Network (CNN) and traditional Gaussian Mixture Model (GMM) are used as the classification models in this study. MGDCC is compared with state-of-the-art magnitude-based features, namely, Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC). In addition, this work also analyzed the noise robustness of MGDCC. To that effect, experiments were performed on various noise types and SNR levels, where the phenomenal performance of MGDCC over other feature sets was reported. Further, this study also analyses the cross-database scenarios for dysarthric severity-level classification. Analysis of Voice onset Time (VOT) and experiments were performed using MGDCC to detect dysarthric speech against normal speech. Further, the performance of MGDCC was then compared with baseline features using precision, recall, and F-1 score and finally, the latency period was analysed for practical deployment of the system.

This work also explores the application of phase-based features on the emotion recognition task and pop noise detection. As technological advancements progress, dependence on machines is inevitable. Therefore, to facilitate effective interaction between humans and machines, it has become crucial to develop proficient techniques for Speech Emotion Recognition (SER). The MGDCC feature

set is compared against MFCC and LFCC features using a CNN classifier and the Leave One Speaker Out technique. Furthermore, due to the ability of MGDCC to capture the information in low-frequency regions and due to the fact that pop noise occurs at lower frequencies, the application of phase-based features on voice liveness detection is performed. The results are obtained from a CNN classifier using the *5*-Fold cross-validation metric and are compared against MFCC and LFCC feature sets.

This work proposed the time averaging-based features in order to understand the amount of information being captured across the temporal axis as there would not be many temporal variations in a cry signal. The research conducted in this study utilizes a *10*-fold stratified cross-validation approach with machine learning classifiers, specifically Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). This work also showcased CQT-based Constant-Q Harmonic coefficient (CQHC) and Constant-Q Pitch coefficients (CQPC) for the classification of infant cry into normal and pathology as an effective representation of the *spectral* and *pitch* components of a spectrum together is not achieved leaving scope for improvement. The results are compared by considering the MFCC, LFCC, and CQCC feature sets as the baseline features using machine learning and deep learning classifiers, such as Convolutional Neural Networks (CNN), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM) with *5*-Fold cross-validation accuracy as the metric.

Keywords: *Infant Cry Analysis, Dysarthria Severity-Level Classification, Emotion Recognition, Voice Liveness Detection, Constant-Q Harmonic Coefficients, Modified Group Delay Function, Noise Robustness.*

List of Principal Symbols and Acronyms

α Greek Alphabet

β Greek Alphabet

Δ Greek Alphabet

γ Greek Alphabet

AESDD Acted Emotional Speech Dynamic Database

AIDS Assessment of Intelligibility and Dysarthric Speech

ASR Automatic Speaker Recognition

CNN Convolutional Neural Network

CQCC Constant-Q Cepstral Coefficients

CQHC Constant-Q Harmonic Coefficients

CQPC Constant-Q Pitch Coefficients

CQT Constant-Q Transform

CV Cross-Validation

EmoDB Emotional Database

FDA Frenchay Dysarthria Assessment

GD Group Delay Function

GDCC Group Delay Cepstral Coefficients

GMM Gaussian Mixture Models

KNN K-Nearest Neighbor

LFCC Linear Frequency Cepstral Coefficients

MFCC Mel Frequency Cepstral Coefficients

MGDCC Modified Group Delay Cepstral Coefficients

MODGF Modified Group Delay Function

POCO Pop Noice Corpus

RF Random Forest

SER Speech Emotion Recognition

SLP Speech Language Pathologist

STFT Short-Time Fourier Transform

SVM Support Vector Machine

UA Universal Access

VLD Voice Liveness Detection

List of Tables

List of Figures

CHAPTER 1
Introduction

Speech is a fundamental mode of human communication that conveys information, emotions, and intentions. Understanding and analyzing speech signals have significant implications across various fields, including healthcare, communication disorders, affective computing, and biometrics. Speech serves as a vital indicator of an individual's well-being, and accurate classification of infant cries plays a crucial role in early childhood healthcare. When it comes to infants, the cry is the only means of communication [26]. Similarly, generating speech requires coordination between the brain and speech production muscles. Lack of coordination results in speech impairments. These speech impairments might be neurogenerative and neurodegenerative [45]. Dysarthria is one such common speech impairment that results in difficulty in speech generation. Furthermore, due to the recent advances in computational ability, voice biometrics are used widely, which also resulted in various spoofing techniques; which in turn motivated the generation of counter-measurement techniques. Additionally, due to these technological advances, dependence on machines is inevitable. Therefore, to facilitate effective interaction between humans and machines, it has become crucial to develop proficient techniques for Speech Emotion Recognition (SER).

1.1 Motivation

Dysarthria is a motor speech disorder resulting from various underlying conditions, such as stroke, traumatic brain injury, Parkinson's disease, or Amyotrophic lateral sclerosis (ALS) [78]. These speech impairments occur as developmental disorders. The inability to reproduce speech causes difficulties for an individual to have effortless communication [6]. This causes individuals to struggle to maintain social relationships and may get prone to mental diseases, such as depression in the later stages. Different types of dysarthria may require specific therapeutic interventions to address the underlying causes and symptoms. Subjective assess-

ment of each individual by a trained speech pathologist can be very expensive and inconsistent. This leads the way to introduce automated dysarthria severity-level classification. By the ability to classify dysarthria into distinct types, clinicians can develop targeted treatment plans that are tailored to the specific needs and challenges of each individual. This personalized approach increases the likelihood of achieving favourable treatment outcomes and improving communication abilities. Additionally, the automated dysarthric classification helps the diagnosis of the areas, which are economically backward and harder to reach for medical professionals. Moreover, the speech production of a dysarthric speaker is significantly different from that of a normal speaker. Due to this, there are a very limited number of assistive technologies for dysarthric speakers. The dysarthria severity-level classification significantly improves the performance of Automatic Speech Recognition (ASR) systems [39].

Another area of problem, where the subject needs careful and proper diagnosis is for an infant cry. For an infant, crying is the only mode through which they can communicate or convey information. However, a cry of an infant can mean many things, it can be a normal cry or a pathological cry. Studies have shown that around 3 million infants die within the first 4 months of birth due to lack of early diagnosis of the disease [3]. Birth asphyxia and sudden infant death syndrome (SIDS) are the leading causes of death for infants [48]. Furthermore, the clinical diagnosis of asphyxia is logistically heavy and time taking and inaccurate sometimes [20]. The acoustic cues of deaf infant cry depend on the type of hearing loss and the age of the pathology detection [61]. However, not every infant is privileged that they get taken care by a group of good paediatricians and receive an early diagnosis. Hence, this encourages the development of an automated infant cry classification system.

In recent times, speech is extensively being used for bio-metric systems, which are linked to social and security [34]. However, the recent advancements in computational systems increase the risk of spoofing attacks on the detection biometric systems [34]. The spoofing attacks occur when an imposter tries to mask as the genuine speaker and access the system. There are many types of spoofing attacks, such as voice conversion systems, synthesizing the original speaker's speech, mimicry, etc [94]. However, among all the attacks, the replay attacks are the most frequently used attack. To that effect, pop noise detection is used as a countermeasure system for the detection of voice liveness. This work proposes a new set of features to improve the performance of the pop noise detection system. Furthermore, the recent advances in technology resulted in a huge de-

pendency on technology. Speech Emotion Recognition (SER) is a key factor in human-computer interaction. By incorporating emotion recognition capabilities into interactive systems, such as virtual assistants, chatbots, and video games, we can create more personalized and responsive user experiences. Emotion-aware systems can adapt their behaviour, responses, and content to better align with users' emotional states, enhancing user satisfaction and engagement. This can lead to improved user experiences and increased usability of technology in various domains.

1.2 Research Problems

1.2.1 Dysarthric Speech Classification

Dysarthria is a disorder represented by difficulties in articulating and pronouncing words due to weak, imprecise, or uncoordinated muscles involved during speech production. It is caused by damage or dysfunction in the central or peripheral nervous system, affecting the muscles responsible for speech production, such as the lips, tongue, vocal folds, and diaphragm. Dysarthria can result from various conditions, including stroke, traumatic brain injury, Parkinson's disease, multiple sclerosis, and certain genetic disorders [24]. Individuals with dysarthria often experience challenges in speaking clearly and intelligibly. Their speech may be slurred, slow, monotonous, or excessively fast, making it difficult for others to understand. Articulation, resonance, phonation, and prosody can all be affected by dysarthria, leading to reduced speech clarity and intelligibility. These difficulties can significantly impact an individual's ability to communicate effectively, affecting their personal relationships, social interactions, and overall quality of life. The usual symptoms of dysarthric speech are [46]:

1. Articulation difficulties
2. Reduced speech intelligibility
3. Impaired prosody
4. Resonance abnormalities
5. Weak voice and less loudness
6. Swallowing difficulties

Dysarthria can be classified into various types depending on the neuromuscular impairment and the area of the speech production system affected. The various types of dysarthria are:

Flaccid Dysarthria

Flaccid dysarthria occurs when there is weakness or paralysis of the muscles involved in speech production. It can be caused by damage to the cranial nerves or the motor neurons in the peripheral nervous system [44]. Symptoms include breathy or hoarse voice quality, imprecise articulation, and reduced loudness. It affects the phonation and respiration of an individual's voice. This results in mispronunciations of consonants.

Spastic Dysarthria

This type of dysarthria is characterized by increased muscle tone and spasticity in the muscles involved in speech production. It can result in slow, effortful speech with strained and tight-sounding articulation due to improper opening and closing of the mouth [100]. Individuals with spastic dysarthria may have difficulty initiating and controlling movements, such as abnormal jaw jerks and facial reflexes.

Ataxic Dysarthria

Ataxic dysarthria is characterized by problems with coordination and control of movements. It is a result of the damage to the core part of the brain that regulates sensory information [38]. Individuals with ataxic dysarthria may exhibit irregular, uncoordinated speech movements, leading to distortion of vowels and consonants.

Hypokinetic Dysarthria

Hypokinetic dysarthria is associated with movement disorders, such as Parkinson's disease. It is characterized by reduced movement, muscle rigidity, and tremors [19]. Speech in hypokinetic dysarthria may be characterized by reduced loudness, monotone or "masked" voice quality, and rapid, repetitive speech patterns. This results in reduced variations of the pitch and loudness.

Hyperkinetic Dysarthria

Hyperkinetic dysarthria is characterized by involuntary movements that affect speech production. It can result from conditions, such as Huntington's disease or certain types of tremors [19]. Symptoms may include excessive or irregular movements of the lips, jaw, tongue, or vocal folds, leading to variable speech intelligibility and control.

Mixed Dysarthria

Mixed dysarthria refers to a combination of dysarthria types. It can occur when multiple areas of the speech production system are affected, such as in cases of neurodegenerative diseases or severe brain injuries [19]. The specific symptoms and characteristics of mixed dysarthria depend on the combination of underlying impairments.

The task of identification of the dysarthric type is performed by pathology experts. Each disorder lies from very low severity to a high severity-level [88]. The analysis of severity or the intelligibility of speech is difficult and might lead to human errors. The severity is determined by the extent of muscle weakness, the degree of impairment in speech intelligibility, and the impact on daily communication activities. Evaluating the severity of dysarthria assists clinicians in developing personalized treatment plans, setting realistic goals, and monitoring progress over time.

1.2.2 Speech Emotion Recognition (SER)

The scientific definition of emotion remains elusive, lacking universal acceptance. It refers to powerful sensations like love or anger, encompassing feelings in general. Emotion is a mental state triggered by neuro-physiological changes, influencing thoughts, feelings, and behaviors. It involves consciousness, bodily sensations, and behaviour, reflecting personal significance. Emotion distinguishes humans from robots and is vital for meaningful human-machine interaction. Researchers classify emotion using four dimensions: duration, quality, intensity, and pleasure. Emotion recognition focuses on identifying emotions, particularly through speech. Speech production involves cognitive processes and physiological aspects of communication. Emotion recognition involves observing visual and auditory cues. This thesis focuses exclusively on analyzing emotions through speech. The gap between human and machine processing hampers accurate identification of a speaker's emotional state. Speech Emotion Recognition

(SER) emerges as a new research field. Effective models depend on understanding the acoustics of various emotions.

1.2.3 Voice Liveness Detection (VLD)

Biometrics refers to the measurement and analysis of unique physical or behavioural traits for identification or authentication purposes. There are various ways of conducting biometric verification. A few of them are face verification [28], iris verification [81], and voice verification [73]. In the realm of biometrics, speech has gained significant attention due to its inherent individuality and the availability of voice-based technologies [34]. Voice liveness, specifically, has emerged as a critical aspect of voice-based biometric systems to ensure the authenticity and security of the captured voice samples. However, with the advancement in technology, the attacks have also increased in numbers. These attacks are known as spoofing attacks. The aim of voice liveness detection is to prevent spoofing attacks, where impostors may attempt to deceive voice-based authentication systems by using prerecorded or synthesized speech. By verifying the liveness of a voice sample, the system can establish the presence of a live human speaker, thus enhancing the security and reliability of voice-based biometric applications [77]. For the same purpose, Automatic Speaker Verification (ASV) systems are designed, which are used to verify specific properties, such as vocal tract system characteristics, pitch features, etc [57]. Some of the attacks performed on the ASV systems are mimicry attacks [30], speech synthesizing [43], conversion of voice [99], and replay attacks [66]. One common attack in replay attacks is recording the original speaker's voice and attempting to fool the verification system. One commonly encountered challenge in voice liveness detection is the absence of "pop noise" in the recorded speech samples. Pop noise refers to the sudden, brief burst of sound at lower frequencies. By detecting and analyzing the presence and characteristics of pop noise in a voice sample, these methods can contribute to the accurate identification of live human speech and the differentiation from synthetic or recorded speech.

1.2.4 Infant Cry Classification

The cry of an infant is a universal language for newborns to convey information. The crying of an infant is a vital and only tool to express their needs, discomforts, and their emotional states [26]. Hence, the cry can be a normal cry or a pathology cry [26]. It becomes difficult for a parent or caretaker to identify if a cry is normal or pathological. Furthermore, the analysis of infant cry holds significant

importance in the early detection and monitoring of certain health conditions or developmental issues. Research has shown that variations in cry acoustics can be indicative of underlying pathologies, such as hearing impairments, neurological disorders, or feeding difficulties. By classifying and analyzing cry signals, healthcare professionals can identify potential problems early on and provide timely interventions and support, optimizing the infant's health and development.

1.3 Social Relevance

Infant cry analysis is a method of automated cry analysis that may assist doctors in the diagnosis of a disease. The social relevance of this research lies in the potential to improve the early detection and diagnosis of health conditions or abnormalities in infants. By accurately classifying cries, healthcare professionals can identify infants, who may require further medical attention or intervention. This early detection can lead to timely treatment, potentially improving outcomes and reducing the impact of certain conditions on the child's development. The experiment has the potential to contribute to the field of pediatric healthcare and enhance the well-being of infants. The practical application of an automated infant cry classifier is through the following ways:

1. **Understanding the cry**: Correctly identifying the reason behind the cry helps to reduce parental stress, while providing the accurate treatment or care necessary.

2. **Developing medical assistive tools**: Automated detection might help in accurately detecting the type of pathology cry. This reduces the chance of delayed treatment in the early stages of diagnosis.

Dysarthria refers to a motor speech disorder that affects the clarity and intelligibility of speech. The experiment aims to classify dysarthria severity-levels. The social relevance of this experiment lies in the potential to improve communication and quality of life for individuals with dysarthria. Accurate classification of severity-levels can assist in personalized treatment planning and speech detection systems. It can help speech-language pathologists plan therapy approaches to meet individual needs. By understanding the severity of dysarthria, appropriate accommodations, and support can be provided to individuals in various social contexts, such as education, employment, and social interactions.

1. **Improving the performance of ASR systems**: The proper classification of dysarthric severity-level helps to improve the ASR system performance by a

large margin. The lack of ASR systems for dysarthric speakers is a very practical problem. For instance, individuals with severe dysarthria may require devices with robust prediction algorithms or text-to-speech capabilities to enhance their communication abilities.

2. **Dysarthric speech enhancement**: Dysarthria severity classification provides insights into the specific speech characteristics and patterns associated with different severity-levels. By understanding these patterns, researchers can develop algorithms or models tailored to enhance speech intelligibility for each severity category. For example, algorithms may focus on reducing vowel distortions, increasing pitch variability, or addressing articulation errors based on the severity-level identified. This work proposes one such feature, phase-based features which can play a crucial role in speech enhancement by using the concept of signal reconstruction.

1.4 Contributions of This Thesis

A novel work has been showcased by applying group delay function-based phase features for the dysarthric severity-level classification task.

- **Phase-Based Features**: The thesis investigates the use of phase-based features for dysarthria severity-level classification. By analyzing the phase information of speech signals, these features capture the irregularities introduced due to the speech disorder. These irregularities are generated due to an increase in production noise, and turbulence in dysarthric speech. It is observed that these are better captured using phase-based features.

- **Noise Robustness of Phase-Based Features**: The thesis investigates the noise robustness of phase-based features. The noise robustness is a very practical problem that is not addressed for the severity-level classification task. By evaluating the performance of these features under various noisy conditions, the thesis provides insights into their effectiveness in real-world environments and their potential for robust speech analysis.

Along with the dysarthria severity classification, given the numerous challenges in infant cry research, an attempt is made to develop an automated infant cry classification system. This is done by proposing the following novel techniques:

- **Time Averaging of Features**: The thesis introduces the concept of time-averaging features for infant cry classification. This novel approach helps

capture variations in infant cries from the spectral-axis alone and shows that the temporal-axis contains the minimum to no information for the cry, providing valuable information for distinguishing between different cry types. These features help to achieve maximum classification accuracy using less computationally complex machine learning classifiers.

- **CQHC and CQPC**: The thesis explores the use of Constant Q-based *harmonic* and *pitch* coefficients as features for infant cry classification by considering the cry signal as a melodic signal. These coefficients capture the fundamental frequency (F_0) and harmonic structure (KF_0, $K \in Z$) of the cry, enabling the discrimination of different cry characteristics related to the infant's needs or discomfort. This showcased the importance of the pitch component for the infant cry classification.

1.5 Organization of Thesis

Fig 1.1 shows the organization of the thesis work as a schematic diagram, which is briefly discussed next:

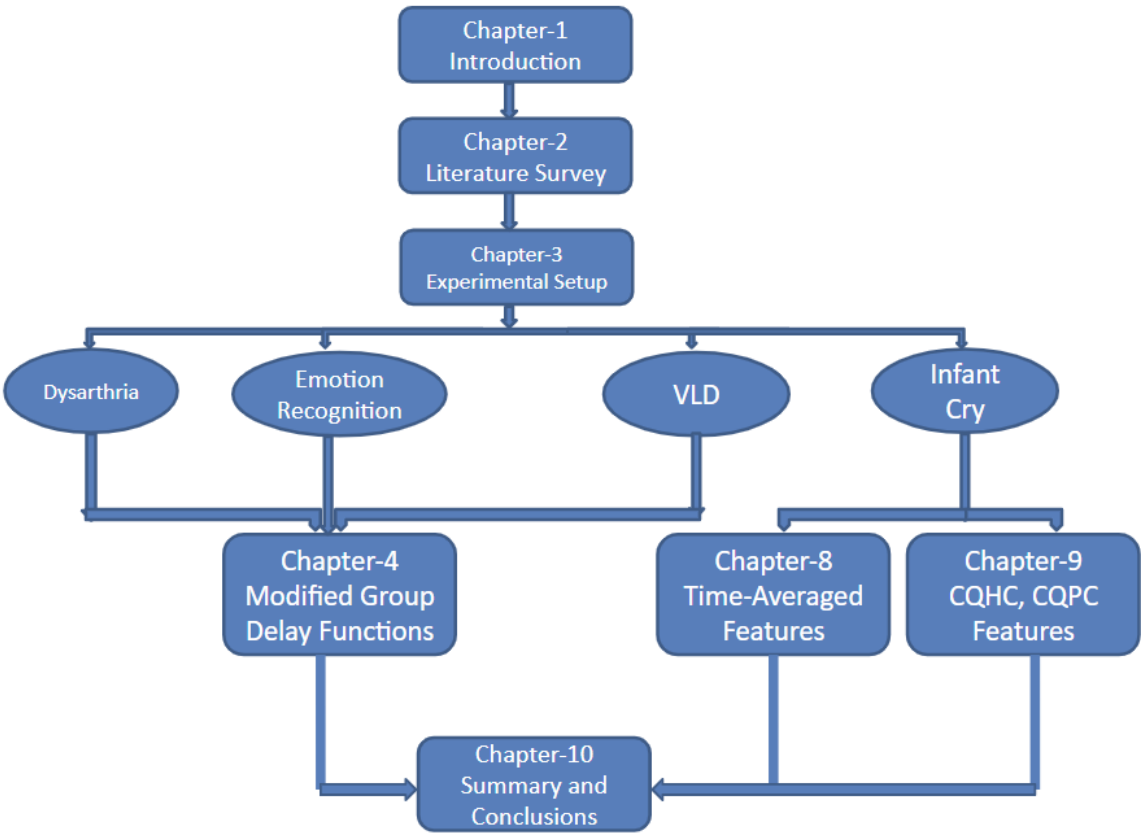

Figure 1.1: Overview of the Organization of Thesis.

- **Chapter 2** presents a detailed study of the previous investigations for infant cry classification, dysarthria severity-level classification, speech emotion recognition, and voice liveness detection. Various methods based on signal processing and deep learning networks on available databases are also discussed.

- **Chapter 3** shows the details of the datasets used in this thesis work, the classifiers used, the baseline features, and the performance metrics for evaluating the models.

- **Chapter 4** presents a detailed explanation of the Fourier transform-based phase features. This is followed by the introduction of group delay function-based features followed by its drawbacks. Later in the chapter, the modified group delay function is explained.

- **Chapter 5** presents the novel approach based on dysarthria severity-level classification using a modified group delay function.

- **Chapter 6** presents the analytical proof of additive noise robust property of group delay features-based function. The analytical explanation is supported by the experimental results performed on dysarthric severity-level classification.

- **Chapter 7 and Chapter 8** shows the application of phase-based features for the emotion recognition task, where the analysis is performed using the Leave One Speaker Out technique. Chapter 8 shows the use of phase-based features on voice liveness detection (i.e., using pop noise detection).

- **Chapter 9** discusses the benefits of averaging features across the time for the infant cry classification. The study has shown that the infant cry signal does not contain much temporal information resulting in the minimal loss, while decreasing the computational cost of the classifiers by a large amount.

- **Chapter 10** proposes a new set of features for infant cry classification, namely, Constant Q Harmonic Coefficients (CQHC) and Constant-Q Pitch Coefficients (CQPC), which are based on the Constant Q Transform (CQT).

- **Chapter 11** concludes the research with an overview of the work completed within the scope of the thesis. Later, the chapter showcases the limitations and future scope of this work.

1.6 Chapter Summary

This chapter gives a brief introduction to infant cry, dysarthria, speech emotion recognition, and voice liveness detection as the problem statements. Later in the chapter, the motivation for this thesis work is explained followed by the social relevance and the contributions of this thesis work. The chapter is concluded by representing the organization of the thesis. In the next chapter, we will see the background and literature on the mentioned problem statements.

CHAPTER 2
Literature Survey

2.1 Introduction

In this chapter, we discuss the literature review of a few studies that have been made in the past for infant cry classification, dysarthric severity-level classification, speech emotion recognition, and voice liveness detection. The chapter starts with infant cry analysis, classification and recent trends followed by dysarthria classification. This chapter also discusses the recent advancements in emotion recognition and finally, discusses voice liveness detection system developments.

2.2 Analysis of Dysarthric Speech

The subject assessment of dysarthric speech requires a diagnosis assessment from Speech Language Pathologist (SLP). The analysis of SLP focuses on the articulation and acoustic parameters of speech signals. There are 4 major methods widely used by SLPs for the assessment of dysarthric speech.

- **Assessment of Intelligibility of Dysarthric Speech (AIDS)**: This method considers the speaking and intelligibility of the speaker. This assessment is performed for speakers above 12 years old. The speech is recorded by the examiner and then played against the panel of judges which rates the speaker on the basis of the intelligibility level of words and sentences [97].
- **Speech Intelligibility Test (SIT)**: It is an electronic form of AIDS introduced in [15]. It provides the score to the examiner. The scoring metrics is the same as the AIDS.
- **Frenchay Dysarthric Assessment (FDA)**: This determines the kind of dysarthria a patient is suffering from [18]. It takes various factors, such as respirations, muscle reflexes, and movement of the jaw, tongue, lips e.t.c into consideration.

- **Dysarthria Profile**: This method takes the facial muscle moments into consideration. The analysis is performed by 1 clinical expert, 1 familiar and 1 unfamiliar listener. This provides a more robust assessment of dysarthric speech [74].

In the literature, the classification of dysarthric speech intelligibility has been approached through two main methods: speech recognition-based techniques and human supervision intelligibility assessment. Various studies have explored different approaches to address this issue. In another study conducted by [36], Mel Frequency Cepstral Coefficients (MFCC) are employed. MFCCs are chosen for their ability to capture the "global" spectral envelope properties, which are relevant in perceptually-motivated audio classification tasks. Additionally, [27] investigate glottal source parameters obtained from a quasi-periodic sampling of vocal tract systems. This approach aims to extract information about the glottal source and its characteristics in dysarthric speech. Moreover, [59] highlight the significance of vocal fold vibration differences or irregularities in dysarthric speech production. They emphasize that these differences cannot be solely characterized by the rate of vibration (i.e., pitch source information), but also by the mode of vibration of the vocal folds.

The traditional magnitude spectral and cepstral features have been used for the classification of the severity-level of dysarthria. In [40], the study shows that measures obtained from fundamental or pitch frequency (F_0) and the second formant frequency (F_2) are highly correlated with the intelligibility of dysarthria. The MFCC showed the ability for speech pathology classification more so for dysarthric speech [6]. In [21], MFCCs are encoded using a deep belief network and used for dysarthria classification using Multi-Layer Perceptron (MLP). Furthermore, the combination of MFCC with auditory features resulted in better results. Later, Linear Frequency Cepstral Coefficients (LFCC) are used to observe the information captured through the linear frequency scale. In [37], LFCC features are used to capture the speech intelligibility of dysarthric speech.

Apart from magnitude spectrum-based features, recent studies showed the importance of phase-based features to improve the performance of speech systems. In [31], the Fourier transform phase-based features are explored for speech and speaker recognition [72], speaker verification in [89], and voice pathology detection [42].

Table 2.1: Available Datasets for Dysarthria Classification Task. After [98], [79], [60].

Database	Speaker	Male/Female	Data
TORGO Database	7	4/3	Words, Sentences
UA-Speech	19	15/4	Words
HomeService	5	3/2	Voice Commands

2.3 Analysis of Emotions of Speech

Initial work on emotion recognition was carried out in late 1999, where Nakatsu R, Tosa N proposed an algorithm for emotion recognition using neural networks. The accuracy obtained was about 50 % [58]. This was then extended by other researchers and now we have multiple emotion recognition features, algorithms, and also datasets are available in various languages [86]. Powerful neural network algorithms are being used to test emotion recognition and accuracy rates have increased ever since [85]. Cognitive features in emotion recognition was also analyzed side-by-side by researchers as their correlation with emotions was found long back [80], [90]. There are three types of emotion databases, namely *acted*, *elicited*, and *simulated* emotions [41], [5]. ***Acted*** emotions refer to emotions that are deliberately portrayed or acted out by individuals, ***Elicited*** emotions are emotions that are intentionally triggered or evoked in individuals through various means. This can be done through external stimuli, such as emotional pictures, videos, or stories, or through interpersonal interactions or specific situations designed to elicit certain emotional responses and ***Natural*** emotions refer to genuine, spontaneous emotional experiences that occur in everyday life situations without any deliberate manipulation or elicitation [85]. The database used in this thesis is of *acted* emotions. Various features are employed for this purpose [17], but the major *four* categories are developed, namely, prosodic, excitation source-based, vocal tract-based, and a combination of the aforementioned features. Prosodic features refer to the suprasegmental aspects of speech that go beyond individual phonemes or words. They involve the rhythm, stress, intonation, and pitch patterns used in spoken language. Prosody plays a crucial role in conveying meaning, emphasis, and emotional expression in communication. Traditional machine learning models, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), K Nearest Neighbour (KNN), Bayes classifier, Support Vector Machine (SVM), random forest, etc were used initially. With the development of Artificial Intelligence (AI) and the emergence of Deep Learning (DL), SER classifications were also shifted to deep learning models. Their ability to automatically

learn complex features, handle large amounts of data, model non-linear relationships, and adapt to diverse tasks and domains make deep learning an increasingly popular and powerful approach in the field of SER. The most commonly used DL models are, Convolutional Neural network (CNN), Recurrent Neural network (RNN), Time Delay Neural Network (TDNN), Long Short Term Memory (LSTM), Residual Neural Network (ResNet), etc

2.4 Voice Liveness Detection

The voice liveness detection in this work is based on Pop Noise detection. The occurrence of pop noise is a natural phenomenon that happens due to bursts of airflow coming through the mouth [70]. There is an inverse relationship between the distance of the speaker and microphone to the energy of pop noise. This relation is used to detect if the speech is genuine or spoofed generated by a replay attack. Voice liveness detection was first proposed in [83], [84]. These studies showcased low-frequency single-channel detection and subtraction-based pop noise detection using 2 channels. In [51], the authors introduced phoneme-based pop noise detection for VLD systems where the duration of pop noise is observed for phonemes. This work classifies the phonemes into 2 categories: Hard and Easy pop noise, respectively. A similar phoneme-based study using Gamma-Tone Cepstral Coefficients (GFCC) was studied in [91]. In [83], the database is proposed pop noise using 3 different microphones. The microphones contain pop noise and no pop noise filters to create a balance in the datasets. The pop noise also gets affected when different kinds of microphones are used. A database named POCO [2] was made available to improve the research in this area. Various works [52], [2] used this dataset and STFT-based features using Deep Neural Networks (DNN), Gaussian Mixture Models (GMM), Support Vector Machine (SVM), and Convolutional Neural Network (CNN) classifiers.

2.5 Analysis of Infant Cry Analysis

The initial works on infant cry analysis started as early as the years of 1960s. The studies performed analysis on 4 pathology cries: pain, hunger, birth, and pleasure [92]. Later, the analysis of cry using a narrow-band spectrogram was started by Q.Xie. et al, in the year 1996 [95]. This work explored 10 different modes of infant cry and studied the pitch and harmonic variations of the infant cry. A new parameter called the H-value is discovered which is found to have a cor-

relation with parent's assessment of infant suffering. In extension to this work, [32] showcased the infant cry analysis on pathological cries. It was noticed that dysphonation and hyper phonation are correlated with the pathology cry modes. Due to recent advancements in technologies and trends, various computer algorithms are employed to analyze the infant cry signal resulting in rapid interpretation and development of analysis tools. The Mel Frequency Cepstral Coefficients (MFCC) which are initially proposed for Western music, have been extensively used for the infant cry classification. The use case of MFCC features for the infant cry classification is proposed in [49], [64]. Later, the effect of the linear filterbank on the infant cry classification was also observed in the studies by using Linear Frequency Cepstral Coefficients (LFCC) features [13], [14]. However, MFCC and LFCC features which are extracted using STFT have fixed resolution in the time-frequency plane. In addition, it fails to preserve *form-invariance* property, as the analysis window used in STFT is a function of *only* time parameter. Hence, study in [65] report the application of CQT-based cepstral features for infant cry classification.

Research in the classification of cry as normal *vs.* pathological has been a recent emerging problem due to its social relevance. The introduction of machine learning and deep learning methods made the automated infant cry classification and analysis faster and more accurate. In [10], the work using Support Vector Machines for classification of normal *vs.* pathological infant cries is reported. The Gaussian Mixture Models (GMM) are also used as classifiers for infant cry classification in [3],[35]. Sudy in [25] shows the application of feed-forward neural networks. Additionally, apart from the traditional classifiers, reports of infant cry classification using acoustic and prosodic features on deep learning architectures, such as CNN, LSTM, and RNN are also proposed. Finally, it is observed that melodic intervals in infant cries are a regular phenomenon indicating it is a healthy cry.

Table 2.2: Available Datasets for Infant Cry Classification Task.

Database	**Creator**	**Recording**	**Source**
Baby Chillanto	NIAOE-CONACYT, Mexico	2268	[75]
Donate a cry	github.com/donateacry	457	[82]
Chatter Baby	chatterbaby.org	1071	[62]
SPLANN	SPLANN Study	13373	[87]
DA-IICT	DAIICT	1190	[11]
ICOPE	infantcope.com	113	[23]

2.6 Chapter Summary

In this chapter, we discussed the attempts made for analyzing and identifying the infant cry, dysarthria, emotion recognition, and voice liveness detection using signal processing techniques with various machine learning and deep learning classifiers. We also analysed different clinical methods for the proposed problem statements. The next chapter discusses the experimental setup followed for this thesis.

CHAPTER 3
Experimental Setup

3.1 Introduction

This chapter describes the experimental setup and the datasets used for dysarthria severity-level classification, infant cry classification, speech emotion recognition, and voice liveness detection. The experimental setup includes the datasets used, classifiers and the statistical metrics used to evaluate this work.

3.2 Database Details

3.2.1 Dysarthria Severity-Level Classification

The severity-level classification is performed on Universal Access Speech Corpus (UA Corpus). The data of 8 speakers (4 male, namely, M01 M05, M07, M09, and 4 female, F02 F03, F04, and F05) are used which can be seen from Table 3.1. From the total of 765-word utterances, a subset of 465 utterances are considered for the feature extraction as mentioned in [29]. Each severity class consists of 936, 920, 936, and 751 samples, respectively. The training data and testing data consist of 80 % and 20 % of the entire dataset, respectively.

Additionally, the results are evaluated on the TORGO dataset. The TORGO dataset consists of speech samples from 7 healthy and 8 dysarthric speakers, respectively. The corpus consists of restricted and unrestricted sentences along with non-words, and words [79]. These consists of English digits, alphabets, the 20 most frequent words in the British national corpus, and 50 words from the Frenchay Dysarthria Assessment (FDA). The recordings are performed using a microphone fixated at the head position at a sampling frequency of 16 kHz. The considered dysarthric diagnosis for this work belongs to speakers with a spastic type of dysarthria. The data consists of 1982 speech samples, where the very low severity-level consists of 671 speech samples, a low severity-level with 627 sam-

ples, and 684 samples belonging to the medium severity-level. In this work, 80 % of the entire dataset is used for training and the rest of 20 % is used for testing. The train test split is done such that both the splits consists of a mix of words and sentences. Table 3.1 shows the speaker statistics of TORGO Corpus.

Table 3.1: Class-Wise Database Details for UA-Speech and TORGO Corpora. After [29], [79].

Severity-Level	UA-Speech	Type	TORGO	Type
Very low	F05, M09	Spastic	F04, M03	Spastic
Low	F04,M05	Mixed, Spastic	F01, M05	Spastic
Medium	F02, M04	Spastic	M01, M04	Spastic
High	F03, M01	Spastic	-	-

3.2.2 Speech Emotion Recognition (SER)

The EmoDB dataset is used. EmoDB is a German speech dataset consisting of 5 male and 5 female actors, whose 10 phrases are recorded. These phrases include 7 emotions, namely, sadness, disgust, anger, neutral, joy, boredom, and fear [9]. The current investigation focused on four emotions, namely, happy, neutral, anger, and sad, with one speaker reserved for the testing data. A total of 383 speech samples are considered. This work adopts the Leave One Speaker Out technique to test the speaker dependency of the feature sets. The train data consists of 338 samples, and the test data consists of 45 samples.

3.2.3 Pop Noise Detection

Voice liveness detection relies on effectively detecting pop noise, which is a crucial factor in the process. To achieve this, the POco COrpus (POCO) is utilized for pop noise detection. This dataset comprises recordings from 66 speakers, with a balanced distribution of 34 male and 32 female speakers. The dataset is carefully designed to cover all 44 phonemes of the English language through the selection of specific words. It is divided into three subsets: RC-A (Recording with Microphone), RP-A (Eavesdropping), and RC-B (Recording with Microphone Array). For this study, the focus is on the RC-A and RP-A subsets when considering the training utterances.

Table 3.2: Statistic of POCO Dataset. After [2].

Data	# Utterance	Male	Female
Train	6952	13	14
Development	3432	6	7
Test	6600	13	13

3.2.4 Infant Cry Classification

To analyze the performance of features for infant cry classification, the Baby Chillanto dataset is used. The recording of this dataset is performed by multiple doctors, who belong to NIAOE-CONACYT, Mexico [76]. The speech signals are sampled at 16 *kHz*. Each cry signal is split into samples of 1-second duration, and it is later grouped into 5 categories. These 5 categories are later clubbed into 2 groups for normal *vs.* pathology classification. The healthy cry signals consist of samples from normal, hunger, and pain. The pathology cry includes samples from asphyxia and dead. Table 3.3 shows the statistic of the dataset. In this work, 80 % of the entire dataset is used for training, and the rest of 20 % is used for testing.

Table 3.3: Statistics of the Baby Chillanto dataset. After [76].

Class	Category	# Utterances
Healthy	Normal	507
	Hungry	350
	Pain	192
Pathology	Asphyxia	340
	Deaf	879

3.3 Classifiers

3.3.1 K- Nearest Neighbors (KNN) Classifier

It operates based on the principle that similar instances tend to belong to the same class. It classifies an input data point based on the majority class among its K nearest neighbors in the feature space. The choice of K, the number of nearest neighbors to consider, is an important parameter in KNN. When the value of K is small, the classifier is more sensitive to noise and individual data points in the dataset. As a result, the decision boundaries can be more complex and irregular, potentially leading to overfitting. As the value of K increases, the influence of individual data points decreases. The decision boundaries become smoother and more generalizable. A larger K value helps in reducing the effect of outliers or

noisy data, leading to better classification accuracy on unseen instances [7]. However, if K becomes too large, the decision boundaries may become overly smooth, potentially leading to underfitting. In such cases, the classifier may oversimplify the data and struggle to capture complex patterns, resulting in reduced accuracy [7].

3.3.2 Support Vector Machines (SVM) Classifier

The SVM is a classification algorithm that aims to find an optimal hyperplane for separating classes, maximizing the margin between support vectors. To achieve this, SVM utilizes kernel functions to transform the data into a higher dimensional space, facilitating linear separation. These kernels can be of various types, such as linear, polynomial, or Radial Basis Functions (RBF).

In addition to the kernel, the SVM classifier also incorporates a regularization parameter known as "c." This parameter plays a crucial role in balancing the trade-off between achieving a larger margin and minimizing training errors. A higher value of c reduces misclassifications but increases the risk of overfitting, while a lower value of c promotes better generalization but may lead to higher training errors.

3.3.3 Random Forest Classifer

The random forest algorithm is an ensemble learning technique that combines multiple decision trees to obtain an output. Each tree is trained on a random subset of the features extracted from the data [7]. The final predictions of the random forest classifiers are made by aggregating the prediction of individual trees. There are a number of parameters to control the random forest classifier. This work fine-tunes parameters that are vital to obtaining a good classification accuracy which are the number of trees, maximum depth of trees, and minimum samples split.

By increasing the number of trees, the classifier can capture a greater variety of patterns and reduce the impact of individual noisy or outlier data points. However, beyond a certain point, adding more trees may not significantly improve accuracy and can increase computational complexity. A higher amount of maximum depth of trees allows you to have more splits and capture intricate patterns among the data. However, increasing the depth beyond a value might lead to overfitting. The Minimum Samples Split determines the minimum number of samples required to perform a split at an internal node or to be considered a leaf

node, respectively. Increasing these values can lead to more robust and generalized trees, reducing overfitting. However, setting them too high may result in underfitting and decreased accuracy, especially when dealing with smaller datasets.

3.3.4 Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a combination of probability density functions that are computed based on a Gaussian assumption. These functions are characterized by mean vectors, covariance matrices, and mixture weights for each component of the mixture. The training of a GMM typically involves the use of the expectation-maximization algorithm, which aims to maximize the likelihood between the classes, as described in [4]. The evaluation of GMM scores is performed using a log-likelihood function.

3.3.5 Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of deep learning model, that employs convolution operation for data processing. It is composed of multiple layers, including convolutional layers, pooling layers, and fully-connected layers. The selection of the number, size, and arrangement of these layers is determined by considering the specific attributes of the data, and the intricacy of the underlying patterns.

Convolutional Layer

The operation of convolution is performed between the input and filter which is also known as *kernel*. These filters or kernels capture the local patterns or spatial relationships, allowing the classifier to learn meaningful features. The number of filters, their size, and the stride (step size) of the convolutions were determined based on the specific requirements of the task. The operation is performed by sliding the kernel through the 2-D input matrix. The stride is the step size in the horizontal direction. A large stride might result in a loss of information. Convolutional operations at the edges of the input can suffer from boundary effects. These effects occur because the filters at the edges have fewer neighbouring pixels to convolve with compared to the filters in the central regions. In order to eliminate this effect, padding is performed.

Pooling Layers

In most cases, pooling layers were added following convolutional layers to decrease the spatial dimensionality and extract crucial features. Pooling plays a role in capturing invariant features and decreasing computational complexity. Max pooling and average pooling are common techniques used for pooling, where the feature maps are downsampled by selecting the maximum or average value within each pooling region, respectively.

Activation Function

To introduce non-linearity into the classifier, activation functions are employed. The Rectified Linear Unit (ReLU) activation function is frequently utilized due to its effectiveness in addressing the vanishing gradient problem.

Dropout Layer

The dropout layer is introduced to avoid overfitting the deep learning model. This helps to improve the generalization ability. This layer drops a percentage of a random set of neurons in the previous layer during the training iterations. By randomly disabling neurons during training, dropout forces the network to learn more robust and distributed representations of the data.

Batch Normalization Layer

The Batch Normalization (BN) layers are introduced for faster convergence of the network. This is based on the concept of normalization of the input makes the network learn faster. In the BN layer, we normalize the activation function of a mini-batch of training samples. This helps to speed up the network training as the normalization causes the loss function to become symmetrical and smooth enabling us to use a larger learning rate without the need to worry about overshooting the minimum point. Additionally, this layer introduces a small amount of regularization due to the randomness injected from the selection of batch samples in random order.

Fully-Connected Layer

The fully-connected layer or dense layers are added at the end of the CNN architecture. The output from the convolutional layers is flattened and fed into the

dense layers. To understand simply, the convolutional layers extract the meaningful, low-dimensional, local features, and the FC layers try to find a non-linear decision boundary to classify the features. This also provides the global view, which is not present in the convolutional layers, i.e., the decision is based on looking at the entire image instead of considering the first few rows/columns of a matrix. For multi-class, the final layer consists of a softmax activation function, and for binary class, it becomes a sigmoid function.

Architecture Details

In this work, The parameters are fine-tuned using the 5-Fold accuracy metric, and it is found that a batch size of 128, a learning rate of 0.001, and epochs of 200 are the best-optimized parameters.

Table 3.4: CNN Architecture for Dysarthria Severity-Level Classification, Emotion Recognition, and Pop Noise Detection

Output Size	**Description**
(20,2000,1)	MGDCC
(20,2000,16)	convolution layer, 16 filters, BN, relu
(10,1000,16)	max-pooling, (2,2), dropout (0.25)
(10,1000,32)	convolution layer, 32 filters, BN, relu
(5,500,32)	max-pooling, (2,2), dropout (0.25)
(5,500,64)	convolution layer, 64 filters, BN, relu
(2,250,64)	max-pooling, (2,2), dropout (0.25)
(2,250,128)	convolution layer, 128 filters, BN, relu
(1,125,128)	max-pooling, (2,2), dropout (0.25)
(1,125,256)	convolution layer, 256 filters, BN, relu
(1,125,256)	dropout (0.25)
128	dense layer, relu
64	dense layer, relu
16	dense layer, relu
4	dense, softmax

CNN classifier is known to learn spatial hierarchies better. Table 3.4 represents the classifier structure used for dysarthria classification, emotion recognition, and voice liveness detection. Table 3.5 indicates the classifier structure used for infant cry classification. Since infant cries and emotion recognition are known to have better spatial information, hence CNN appears to be a better choice of classifier than the other classifier structures.

Table 3.5: CNN Architecture for Infant Cry Classification

Output Size	Description
(20,130,1)	CQHC
(20,130,16)	convolution layer, 16 filters, BN, relu
(10,65,16)	max-pooling, (2,2), dropout (0.25)
(10,65,32)	convolution layer, 32 filters, BN, relu
(5,32,32)	max-pooling, (2,2), dropout (0.25)
(5,32,64)	convolution layer, 64 filters, BN, relu
(2,16,64)	max-pooling, (2,2), dropout (0.25)
(2,16,16)	convolution layer, 16 filters, BN, relu
(2,16,16)	dropout (0.25)
(2,16,16)	convolution layer, 16 filters, BN, relu
(2,16,16)	dropout (0.25), followed by flattening
128	dense layer, relu
64	dense layer, relu
64	dropout (0.25)
1	dense, sigmoid

3.4 Performance Evaluation Metric

3.4.1 K-Fold Cross Validation Technique

The end goal of machine learning and deep learning models is a good generalization. The cross-validation setup helps us to estimate the model's ability to generalize. This work uses the K-Fold cross-validation technique. Here the dataset is split into K fold, where the K-1 fold is used for testing and the rest of the folds are used as training data. The K represents a number of splits of the dataset. The K-Fold also provides confidence in the score values as the test score of the K-Fold is a result of the average of scores from K models. The stratified technique works in a way such that each fold is a good representative of the entire dataset. This helps us to avoid any kind of data imbalance.

3.4.2 Confusion Matrix

The confusion matrix serves as a technique to summarize the performance of a classifier. It provides a more comprehensive understanding of the accuracy of the classification model by indicating both correct classifications and the types of errors being made. The combinations of actual and predicted values in a confusion matrix result in the following parameters:

- **True Positive (TP)**: The positive values are correctly predicted as positive.
- **False Positive (FP)**: The negative values are incorrectly predicted as positive.
- **False Negative (FN)**: The positive values were incorrectly predicted as negative.
- **True Negative (TN)**: The negative values were correctly predicted as negative.

3.4.3 Accuracy

It is one of the most simplest and powerful classification metrics. Accuracy provides a fair evaluation of the model only when the dataset is adequately balanced. The accuracy score is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.1}$$

3.4.4 Precision

It is a measure of how many test samples predicted were correctly predicted for a class. It gives us information about how precisely the model has made predictions for a class. This metric is important when the cost of False Positive (FP) is higher. Precision is calculated as:

$$Precision = \frac{TP}{TP + FP}. \tag{3.2}$$

3.4.5 Recall

It is a metric that quantifies the accuracy of correctly predicting samples belonging to a specific class. It measures the ratio of correctly predicted positive outcomes to the total number of positive outcomes. This metric is particularly relevant when the cost of false negatives (FN) is significant. The recall metric is calculated as:

$$Recall = \frac{TP}{TP + FN}. \tag{3.3}$$

3.4.6 F1-Score

It is a powerful metric as it can be used for both balanced and unbalanced data. This metric provides a balance between the precision and recall of a classification

model. Higher the F1-Score, the better the classifier. The F1-Score is calculated as the harmonic mean of precision and recall, and it can be calculated as:

$$F1 - Score = 2.\frac{Precision.Recall}{Precision + Recall}. \tag{3.4}$$

3.4.7 Area Under Curve (AUC)

AUC is an important metric for imbalanced multiclass dataset classification. The AUC curve measures the separability between the classes. The higher AUC, the better the model. The AUC depends on 2 metrics, False Positive Rate (FPR) and True Positive Rate (TPR). The TPR metric is the same as recall and the FPR metric is calculated as:

$$FPR = \frac{FP}{TN + FP} \tag{3.5}$$

3.5 Baseline Features Used

3.5.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a well-known feature set used for various applications. To extract the MFCC features, the signal is pre-emphasized to amplify the energy content in the higher frequencies. This is followed by windowing the speech signal into short instances to make the speech signal a stationary signal. To avoid the generation of high-frequency noise and to maintain the continuity of signal even after windowing, we use an overlap length along with the window length. Later, the signal is transformed into the frequency domain and a Mel filterbank is used to change the frequency scale to the Mel frequency scale. If s indicates the frequency of the signal, the conversion into the Mel scale is done as follows [33]:

$$Mel(s) = 2595 * log(1 + s/700).$$

The magnitudes of the power spectrum obtained from the filterbank are passed through the logarithmic operator. Further, a DCT function is applied to the logarithmically compressed filterbank outputs to obtain a compact representation of the spectrum. The resulting coefficients are called MFCC.

The MFCC features are initially generated for musical signals as they capture the pitch, and timbre information well. Since this work considers the infant cry signal as a melodic signal, this becomes a solid baseline feature to outperform. It should also be noted that the pitch information of a human speech signal plays a

vital role in speech emotion recognition as the pitch varies drastically from emotion to emotion. Additionally, this feature is also used as a baseline for dysarthria and pop noise detection due to its resolution at low frequencies. It is observed that the natural production noise of the speech production system increases with voice disorder, and this production noise is present at lower frequencies. Furthermore, pop noise is a short-duration acoustic disturbance that is present at lower frequencies, that often occurs during the production or recording of live speech.

3.5.2 Linear Frequency Cepstral Coefficients (LFCC)

The LFCC feature extraction technique is quite similar to that of the MFCC feature extraction. The only change lies in the filterbank, we replace the Mel filterbank with a linear filterbank.

The pitch of an infant is higher than that of an adult male or female speech. Since the pitch lies in the higher frequencies for a cry signal, a better resolution at higher frequencies might capture important information. To explore this and given the poor resolution of MFCC features at high frequencies, LFCC features are considered as the baseline features for the infant cry classification system. The same applies to emotion recognition. Additionally, it is used as a baseline feature in dysarthric severity classification in order to observe the increase in the linearity of the formant structure with the dysarthric severity-level.

3.5.3 Constant-Q Cepstral Coefficients (CQCC)

The CQT-based cepstral coefficients have the ability to vary the spectro-temporal resolution as it has a window function that is dependent on both time and frequency. Additionally, this transform poses the form-invariance property, which is used in the spectral-domain for pattern classification. Adding to this, the initial study of CQT [8] indicates that this transform is designed to improve the note resolution of music. Since this study considers the infant cry signal as a melodic signal, this becomes a good baseline feature for the study. The CQCC features are obtained by passing the CQT-based spectral coefficients through the DCT block.

3.6 Chapter Summary

This chapter discussed the datasets used for various problem statements, such as dysarthria severity classification, infant cry classification, emotion recognition, and voice liveness detection. Later, this chapter describes the classifier details

used in this work and the performance evaluation metrics used for the classification systems. In the next chapter, we present the group delay function and its properties.

CHAPTER 4
Group Delay and Modified Group Delay-Based Features

4.1 Introduction

The speech signal is represented by the features obtained from the Short-Time Fourier Transform (STFT) function, which yields a magnitude and phase-based representation of the signal. While the significance of phase in speech signals has been recognized in various studies, incorporating the phase spectrum into speech applications remains challenging due to its complexity. One approach to acquiring Fourier transform-based phase features is through the *Group Delay* (GD) function. However, the GD function has certain limitations, leading to the development of the *Modified Group Delay Function* (MODGF). This chapter explores the Fourier transform-based phase spectrum, discusses the properties of the Fourier phase spectrum, introduces the group delay function and its properties, and finally presents the modified group delay function as a solution to the drawbacks of the group delay function.

4.2 Fourier Transform Phase

It is well known that a speech signal is a quasi-periodic, non-stationary signal that continuously changes w.r.t time. As the human auditory system processes the speech signal w.r.t frequency bands, the STFT analysis is performed to capture the information of a speech signal. To define the speech signal completely, both magnitude and phase components are necessary since the major speech production systems, such as the vocal tract system belong to the minimum phase and the glottal source belongs to the maximum phase system in particular, both zeros and poles of the glottal transfer function lies outside the unit circle. Initially, studies believed that the human auditory systems are *phase deaf* [50]. Later,

studies have proven that distorting the phase makes the speech signal unrecognizable [1]. Even though the importance of phase-based features for speech processing has been studied, there has been a minimum effort for the application of phase-based features for applications such as dysarthria severity classification, emotion recognition and voice biometrics. It is also observed that the phase-based features capture the irregularities in the speech signal such as turbulence which helps to categorize the speech [16]. Consider the equation x(t)=1+ $0.5cos(2\pi t+\phi_1)$ $+cos(4\pi t+\phi_2)+ 0.66\ cos(6\pi t+\phi_3)$. From Figure 4.1, it can be noticed that changing the phase can change the signal. However, this cannot be observed in the magnitude spectrum. Hence, since the speech production system is a non-linear phase system and in order to observe the phase distortions which might be important for intelligibility assessment, one needs to observe the phase spectrum.

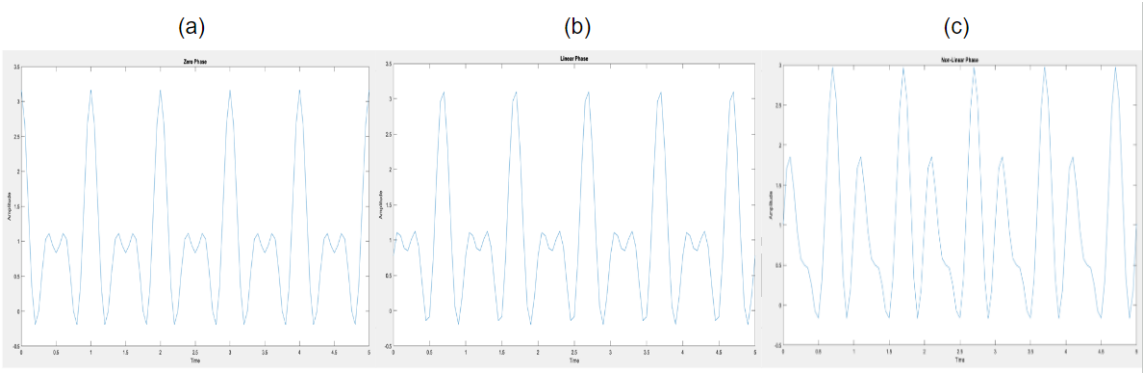

Figure 4.1: Plots for Signal x(t) at Various Phase Angles. (a) Zero Phase,(b) Linear Phase, and (c) Non-Linear Phase where, x-axis indicated time and y-axis indicates amplitude of the signal.

4.2.1 Properties of Fourier Transform Magnitude Spectrum

1. If *x(n)* is a *real* signal, then the magnitude spectrum is an *even* function of ω.

2. When the Inverse Fourier Transform is applied to a magnitude spectrum, it produces a zero-phase signal. This implies that the magnitude spectrum is independent of the phase of the original speech signal.

3. If the impulse response of a signal is of a cascade of resonators, then the magnitude spectrum of the signal is given by the product of the magnitude spectrum of each resonator.

4.2.2 Properties of Fourier Transform Phase Spectrum

1. For the signal x(n) which is a *real* signal, the phase spectrum is an *odd* function of ω.
2. If the signal is shifted in the time-domain, it is translated as a linear phase component in the phase-domain.
3. The Inverse Fourier Transform of the phase-spectrum is an all-pass signal.
4. If the signal x(n) is a cascade of resonators, the phase spectrum is the sum of the unwrapped phase of each resonator.
5. The phase unwrapping technique is non-trivial indicating that there is no specific way for unwrapping the wrapped phase.

4.3 Group Delay Function

Identifying resonance frequencies or formant frequencies from the phase spectrum poses a challenge because they are obscured by the phase wrapping phenomenon at multiples of 2π. To overcome this issue, the signal must be a minimum phase signal, where the continuous (i.e., unwrapped phase) phase is denoted by $\theta(e^{j\omega})$. Minimum phase signals are preferred because their magnitude spectrum and group delay spectrum exhibit similar characteristics. The group delay function, which is the negative derivative of the unwrapped Fourier transform phase, is used to quantify this relationship. The group delay function is defined as [55]:

$$T(e^{j\omega}) = -\frac{d\theta(e^{j\omega})}{d\omega} \tag{4.1}$$

The group delay function can directly be derived from the signal z(n) as follows:

$$T_z(e^{j\omega}) = -Im(\frac{dlog(Z(e^{j\omega}))}{d\omega}) = \frac{Z_R V_R + Z_I V_I}{|Z(e^{j\omega})|^2} \tag{4.2}$$

where z(n) is the signal and v(n) is nz(n). Since the magnitude spectrum is an even function, the log magnitude spectrum can be given as [53]:

$$ln|P(e^{j\omega})| = \frac{c(0)}{2} + \sum_{n=1}^{\infty} c(n)cosn\omega \tag{4.3}$$

The phase spectrum is an odd function. Hence, the unwrapped phase is expressed as:

$$\theta(e^{j\omega}) = -\sum_{n=1}^{\infty} c(n) sinn\omega. \tag{4.4}$$

From equation 4.4, the group delay is obtained by calculating the negative derivative of the unwrapped phase.

$$T(e^{j\omega}) = \sum_{n=1}^{\infty} nc(n) cosn\omega. \tag{4.5}$$

From eq (4.3), (4.4), and (4.5), it is observed that the magnitude-based features are related to the phase-based and group delay features through the cepstral coefficients. It can also be observed that the group delay function is calculated by weighted cepstrum. While for maximum phase systems, the eq (4.4) and (4.5) become as follows [96]:

$$\theta(e^{j\omega}) = \sum_{n=1}^{\infty} c(n) sinn\omega, \tag{4.6}$$

$$T(e^{j\omega}) = -\sum_{n=1}^{\infty} nc(n) cosn\omega. \tag{4.7}$$

4.3.1 Properties of Group Delay Function

The properties of the group delay are stated below:

1. The presence of poles and zeros in the transfer function is manifested as peaks and valleys in the group delay spectrum, respectively.

2. Convolution of the time domain signal results in additivity in group delay domain signal. Let $x(n) = x_1(n) * x_2(n)$, the magnitude spectrum results in $X(e^{j\omega}) = X_1(e^{j\omega})X_2(e^{j\omega})$ where each $X_i(e^{j\omega})$ is a response of an individual resonator. The magnitude response is given as:

$$|X(e^{j\omega})| = |X_1(e^{j\omega})||X_2(e^{j\omega})|. \tag{4.8}$$

   The phase response and the group delay response is :

$$argX(e^{j\omega}) = argX_1(e^{j\omega}) + argX_2(e^{j\omega}), \tag{4.9}$$

$$T_X(e^{j\omega}) = -\frac{d(argX_1(e^{j\omega}))}{d\omega} - \frac{d(argX_2(e^{j\omega}))}{d\omega}. \tag{4.10}$$

$$T_X(e^{j\omega}) = T_{X1}(e^{j\omega}) + T_{X2}(e^{j\omega}) \tag{4.11}$$

3. The high resolution characteristic of the group delay function leads to improved resolution in identifying closely spaced resonant peaks or formant peaks.

4.3.2 Issues with Group Delay Spectrum

The group delay function is specifically applicable to minimum phase signals [56], while the speech signal is a mixed-phase system that contains zeros introduced by noise or nasal sounds. These zeros, located near the unit circle, are manifested as prominent spikes in the group delay spectrum. In 4.2, two zeros α and β are depicted. Based on triangle properties, the largest interior angle corresponds to the side opposite it. Consequently, the chord between ω_1 and ω_2 will be longer for zeros in proximity to the unit circle. This implies a higher rate of change for zeros near the unit circle. Furthermore, larger angles result in shorter distances to frequency bins and smaller denominators in the group delay function. These undesired spikes hinder the identification of original formant locations and cannot be effectively eliminated through smoothing techniques. These spikes emerge due to reduced denominators in Equation 4.2, leading to larger group delay values. To address these undesirable spikes, the Modified Group Delay Function (MODGF) is introduced [54].

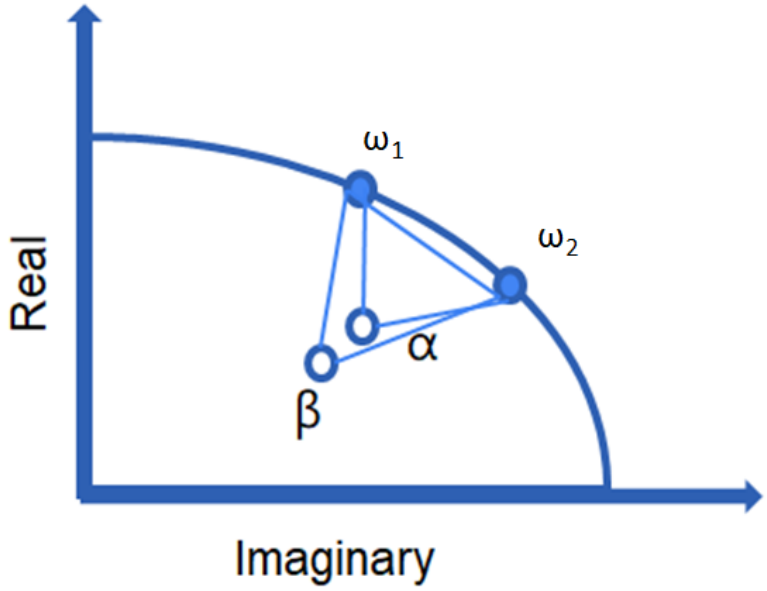

Figure 4.2: Issue with GD. ω_1 and ω_2 represents the adjacent frequency bins.

4.4 Modified Group Delay Function (MODGF)

Voiced speech comprises both causal and non-causal components. In order to address the issue of unwanted spikes and achieve a meaningful representation of the group delay spectrum, the Modified Group Delay Function (MODGF) is introduced. When the zeros are positioned near the unit circle, the denominator of the group delay function becomes very small, resulting in a spiky and limited dynamic range of the group delay for a speech signal. To mitigate this spiky nature, the denominator of the group delay function is replaced with a centrally-smoothed version, which represents the envelope of the source information. The modified group delay function effectively moves the zeros radially inside the unit circle, thereby reducing the occurrence of spikes in the valleys. Let the frequency domain representation of g(n) and ng(n) be G(ω) and H(ω). The ng(n) helps to generate the delayed signal. A cepstrally smoothened |G(ω)| is computed. The cepstral smoothened signal helps to restore the dynamic range, and reduce the spiky structure of the phase-based features. Later, the modified group delay function is computed as:

$$T_m(\omega) = \frac{T(\omega)}{|T(\omega)|}|T(\omega)|^{\alpha}, \tag{4.12}$$

where,

$$T(\omega) = \frac{G_R(\omega)H_R(\omega) + G_I(\omega)H_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{4.13}$$

where $S(\omega)$ represents the cepstrally smoothed version of $G(\omega)$ and G_R, G_I, H_R, and H_I indicate the real and imaginary parts of $g(n)$ and $ng(n)$, respectively. Two new parameters, α and γ are introduced, which are used to restore the dynamic range and reduce the amplitude of the unwanted spikes. α and γ lies between $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$.

4.5 Chapter Summary

This chapter discussed the brief technical details and the application for speech applications of Fourier transform-based phase features. The importance of phase features and their properties. Later, the group delay and modified group delay functions are introduced. The next chapter shows the application of these phase-based features for dysarthria severity-level classification.

CHAPTER 5
Modified Group Delay Features for Dysarthric Severity Classification

5.1 Introduction

Dysarthria is a neuro-motor speech disability that impairs speech comprehension, however, it is typically undetectable to humans. In addition to assisting automatic dysarthric speech recognition systems, dysarthric speech severity-level classification serves as a diagnostic tool for assessing the progression of a patient's severe condition and recognition of dysarthric speech-an important assistive speech technology. The phase-based features are used for voice pathology detection due to their ability to capture the irregularities caused due to disordered speech. The dysarthria severity causes an increase in turbulence leading to introduce higher irregularities than normal speech. The phase information is known to have more temporal (and also transitional) information of the speech signal than the magnitude spectrum information. This study investigates the effect of phase-based features over magnitude-based features for dysarthric severity-level classification. To that effect, Modified Group Delay Cepstral Coefficients (MGDCC) are imposed against state-of-the-art features, namely, Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) using CNN and GMM as the classifiers with accuracy, precision, recall, and F1-score as performance metrics on UA-Speech and TOROG datasets. Additionally, the effect of dynamic features for dysarthria severity classification is also observed. Furthermore, to verify the speaker independency of the proposed feature set, cross-database analysis is performed using the CNN classifier. Finally, to understand the practicality of the proposed feature set, latency period analysis is performed. From the results, it is seen that the proposed features surpass the baseline features comfortably. The experiments are performed by splitting the data into 80 % training and 20 % testing set.

5.2 Proposed Features

This work proposes modified group delay-based features for dysarthria classification. The feature extraction procedure involves computing the STFT of the signals x(n) and nx(n), respectively. Let the frequency domain representation of g(n) and ng(n) be G(ω) and H(ω). The ng(n) helps to generate the delayed signal. A cepstrally smoothened $|G(\omega)|$ is computed. The cepstral smoothened signal helps to restore the dynamic range and reduce the spiky structure of the phase-based features. Later, the modified group delay function is computed as:

$$T_m(\omega) = \frac{T(\omega)}{|T(\omega)|}|T(\omega)|^{\alpha}, \tag{5.1}$$

where,

$$T(\omega) = \frac{G_R(\omega)H_R(\omega) + G_I(\omega)H_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{5.2}$$

where $S(\omega)$ represents the cepstrally smoothed version of $G(\omega)$.

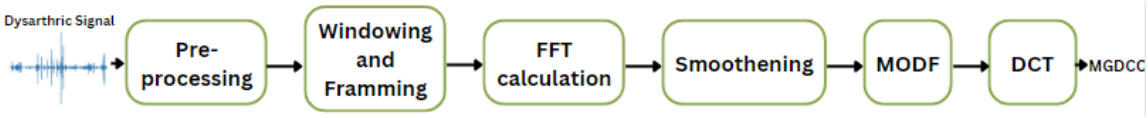

Figure 5.1: Functional Block of Proposed MGDCC Feature Extraction for the Dysarthric Severity-Level Classification System.

In order to obtain the cepstral features of the modified group delay function, a DCT operation is performed. The first coefficient of the DCT is neglected in this work. The first coefficient value indicated the average value in the group delay function. Due to the factors like linear phase resulting due to window and location of the pitch w.r.t window, the importance of the DC coefficient is an unexplored area. The functional block diagram for the feature extraction procedure is shown in 5.1.

5.3 Motivation of Phase-Based Features for Dysarthria

Studies have shown that the group delay features capture the irregularities and the turbulences produced during the phonation [16]. To study the variation of production noise for a dysarthric speaker wrt severity level, an acoustic parameter known as Voice Onset Time (VOT) is studied. The speech production system generates production noise due to variations occurring from various sources, such

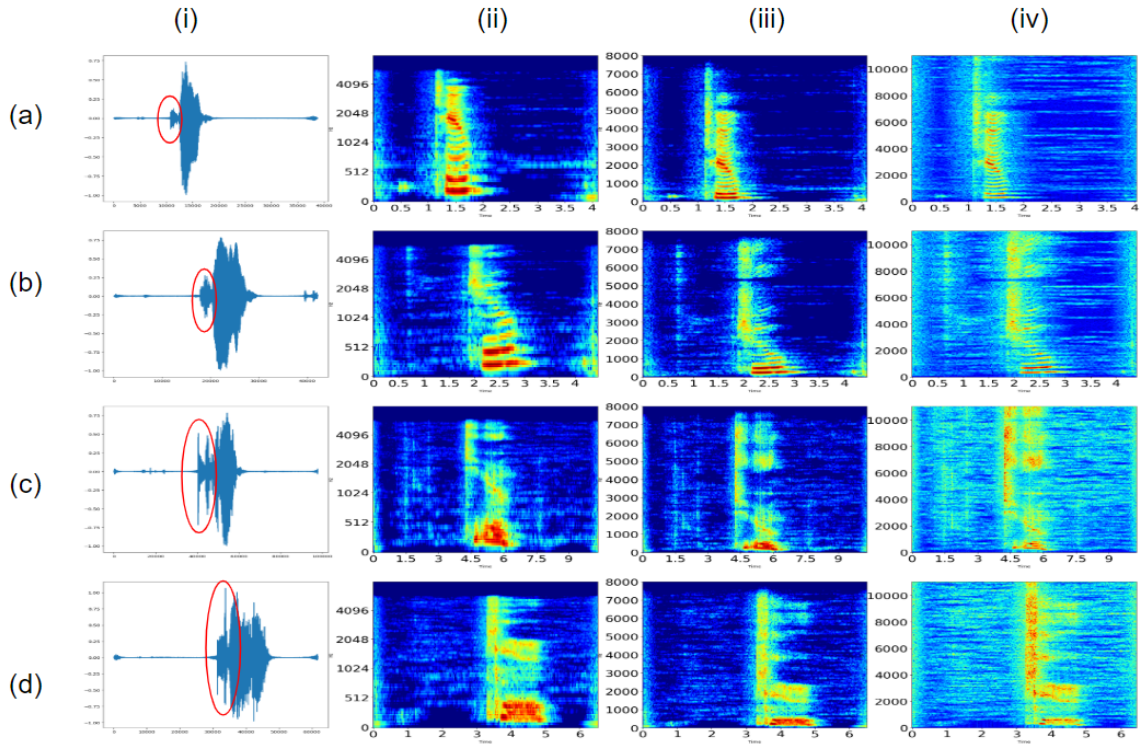

Figure 5.2: Fig. 5.2(i), Fig. 5.2(ii), Fig. 5.2(iii), and Fig. 5.2(iv) of Each Panel Depicts the Time-Domain Waveform, Mel Spectrogram, STFT Spectrogram, and Modified Group Delay-Gram of the Clean Dysarthric Speech Signal of the Word *"to"*. Fig. 5.2(a), Fig. 5.2(b), Fig. 5.2(c), Fig. 5.2(d) Indicate Very Low, Low, Medium, and High Dysarthric Severity-Level Clean Speech Signals, respectively. The Voice Onset Time (VOT) Regions of Various Dysarthria Severity-Levels are Circled. Best Viewed in Colour.

as the vocal tract systems. This production noise contributes to the generation of various speech sounds, such as plosive and fricative consonants. The VOT is an acoustic property that captures the vocal fold vibrations during the release of plosive consonants. This acoustic property provides insight into the production of stop consonants, and how it gets affected by dysarthria severity-level. It can be observed from Fig. 5.2, the time-domain plots show that along with the duration, the energy of the VOT region also increases with the severity-level. It might be because of the lack of speech system muscle coordination as the tongue moves towards the lips. Additionally, it is observed that the duration of VOT for a very low severity-level is 124 ms, while it is 252 ms for a high severity-level speaker. It can be observed that MGDCC features are able to capture this low frequency VOT information, whereas the magnitude-based features fail to capture the VOT information.

5.4 Experimental Results

5.4.1 Parameter Tuning for MGDCC

MGDCC involves two constraint parameters, alpha (α), and gamma (γ). The parameters are fine-tuned using a greedy search algorithm, i.e., the parameters are optimized wrt to the performance of severity-level classification. The parameters are varied from 0 to 1 with a step size of 0.1. The evaluation is done based on the CV scores of the CNN classifier. The best fold accuracy is achieved for α **= 0.1** and γ **= 0.3**, resulting in an accuracy of **96.53** % and **96.46** % for UA-Speech and TORGO corpora, respectively. This result indicates that the obtained alpha and gamma parameters are the generalized parameters for the dysarthria severity-level task. The effect of parameters can be seen in Fig 5.3.

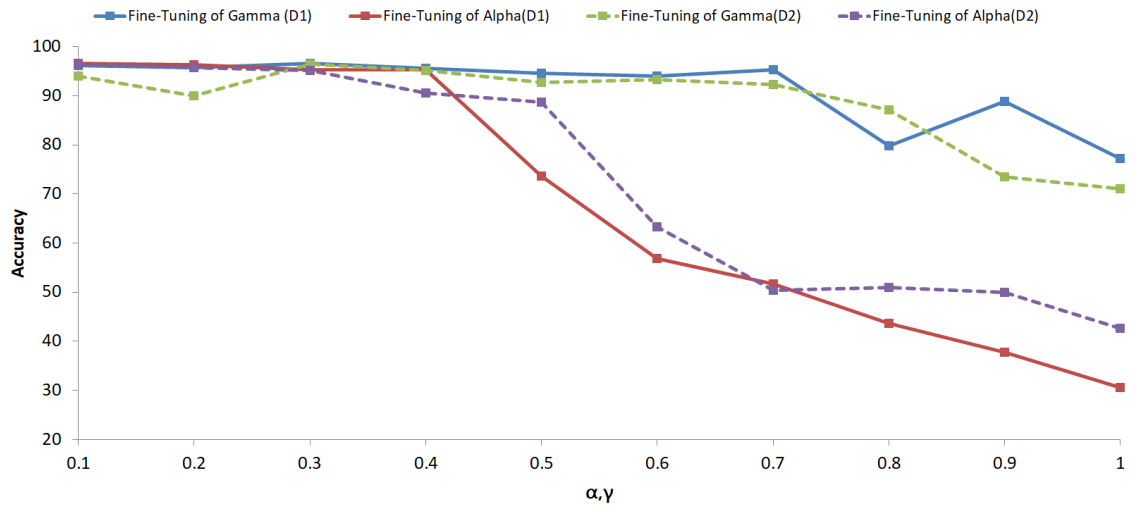

Figure 5.3: Fine Tuning of α and γ Using Greedy Search Technique for Dysarthric Severity-Level Classification on UA-Speech Corpus (D1), and TORGO Database (D2).

5.4.2 Dimensionality Tuning for MGDCC

The dimensions of the MGDCC features are varied from a standard size of 13 to 30 with a step size of 3. The *20*-D features resulted in the best accuracy of 96.53 % when compared with the other dimensions.

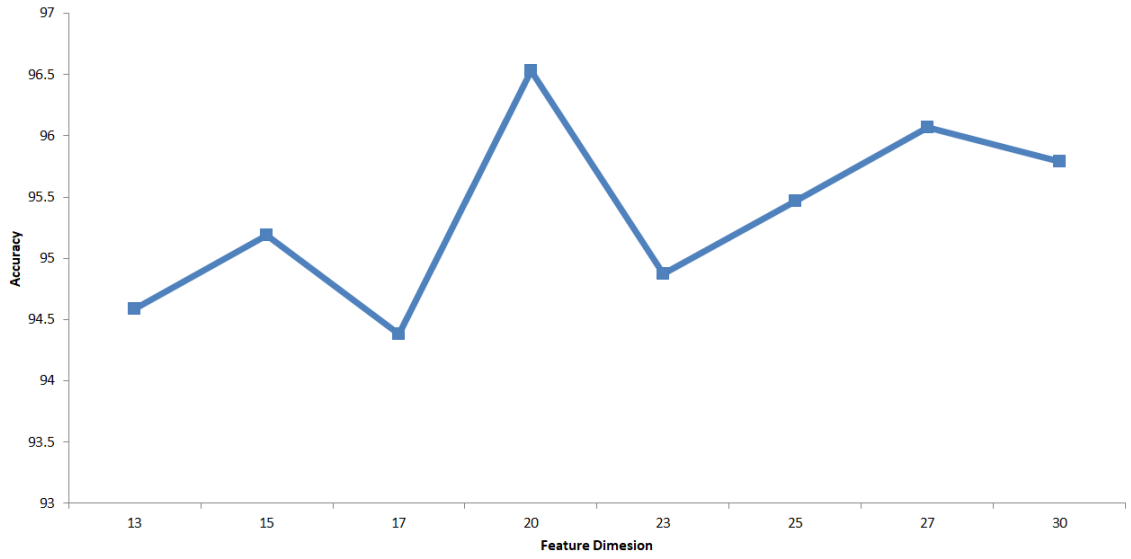

Figure 5.4: Fine Tuning of Feature Dimension on UA-Corpus.

5.4.3 Results for Dysarthric Severity Classification

Experiments are performed using stratified *5-fold* CV. The results are performed on Convolutional Neural Networks and simple machine learning classifiers, such as Gaussian Mixture Models. CNN classifier is known to learn spatial hierarchies better, which are vital differentiating features in the speech signal and hence, CNN appears to be a better choice of classifier than the other classifier structures. From Table 5.1, it can be observed that the cepstral coefficients with a linear filterbank (i.e., LFCC) result in a better classification accuracy than the Mel filterbank (i.e., MFCC). This indicates that the linearity of the dysarthric speech increases as the severity-level increases resulting in more information capture using a linear filterbank. It can also be seen that the phase-based features (i.e., MGDCC) achieved a higher classification accuracy. when compared to baseline magnitude-based MFCC and LFCC features (with an absolute improvement of **5.23 %** and **2.40 %**, respectively). Hence, the phase-based cepstral features capture crucial information that aids in the classification of dysarthric severity-level. Since MGDCC returns higher precision, the number of false positives is less compared to the baseline features, and higher recall signifies that the count of false negatives is less. Hence, the accuracy results can be supported by the F1-score metric.

On the contrary, the CNN classifier performs better than the traditional GMM classifier. This indicates that the GMM classifier is not suitable for the severity-level classification of dysarthria. Furthermore, GMM is only based on mean and variance, which might not be adequate to capture the non-linearity patterns in the speech production mechanism and dysarthric speech. The inclusion of higher-

order moments requires training data with longer-duration dysarthria speech data. On the other hand, the proposed deep learning architecture, CNN is able to capture the non-linearities in speech patterns from short-duration speech signals. However, the performance of the GMM classifier improves in the TORGO database as the database contains longer-duration sentences along with the words. This enables GMM to capture the patterns better when compared to the short-duration speech that is present in the UA-Speech corpus.

Table 5.1 also shows that traditional group delay function features do not fare well for dysarthric severity-level classification. This is due to large spikes produced by mixed-phase signals, such as speech. These large spikes in the group delay spectrum *mask* the formant information, making the features least informative for the classifier. In particular, as per Manfred Schroeder, "Human being emits and perceive sounds by emitting spectral peaks (resonances) and the spectral valleys (anti-resonance)." Thus, the GDF disrupts the formant structure and hence, affects the perception of sounds and thereby, speech intelligibility.

Table 5.1: % Classification Accuracy for Various Feature Sets using GMM and CNN Classifiers on UA-Speech (D1) and TORGO (D2) Corpora. The values in the brackets indicate the test accuracy.

Dataset	Classifier	Features	5-fold	Precision	Recall	F1-Score
D1	GMM	MFCC	84.14	80.22	83	81.58
		LFCC	84.14	81.97	83.16	82.56
		GDCC	73	71.19	70.03	70.60
		MGDCC	85.48	85	83.21	84.09
	CNN	MFCC	91.30 (91.49)	92.36	93.04	92.07
		LFCC	94.13 (94.90)	93.94	93.98	93.95
		GDCC	73.60 (71.24)	71.04	75.72	73.26
		MGDCC	**96.53 (96.75)**	**96.71**	**96.76**	**96.52**
D2	GMM	MFCC	83.92	74.55	70.48	72.40
		LFCC	83.21	72	76.70	74.27
		GDCC	71.86	73.11	75.95	74.48
		MGDCC	**83.92**	**74.32**	**75.18**	**74.74**
	CNN	MFCC	87.67 (85.64)	89.67	87.89	88.77
		LFCC	92.49 (92.41)	**94.37**	92.94	93.64
		GDCC	74.02 (74)	84.91	76.03	74.29
		MGDCC	**96.46 (96.71)**	94	**94.12**	**94.05**

5.4.4 Effect of Dynamic Features

The study of the effect of dynamic features is studied on UA-Speech Corpus. The experiments are only performed on the CNN classifier due to its ability to capture the non-linearities and due to the poor performance of the GMM classifier as observed in the previous section. It is observed that the average time duration of very low, low, medium and high severity-levels is 2.249, 2.536, 3.436, and 4.316 seconds, respectively. Hence, as the severity-level increases, the speech variations

Table 5.2: Effect of Dynamic Features on UA-Corpus Dataset Using CNN Classifier. The values in the brackets indicate test accuracy.

Features	Δ	$\Delta\Delta$
MFCC	73.39(75.67)	64.20(60.19)
LFCC	75.51(78.41)	62.33(65.92)
GDCC	22.57(20.35)	21.90(20.36)
MGDCC	**85.33 (86.52)**	**65.12(69.60)**

for a shorter time instance decrease. This statement can be proved experimentally by considering the delta (Δ) and delta-delta ($\Delta\Delta$) features. Since the variations in the speech signal approach are constant as the severity-level increases, the information captured using the dynamic parameters decreases as the rate of change almost becomes 0. This causes low accuracy values, when compared with the static features.

5.4.5 Cross-Database Evaluation

Further, experiments of cross-database are performed between 2 datasets (UA-Speech and TORGO) using CNN classifier. Both the databases are re-sampled to 22.050 *kHz*. The classifiers till now might be able to take advantage of the similar dysarthric type and uniformity of the data present in train and test split. The cross-database evaluation helps to confirm if the feature is actually able to capture the dysarthric-based features. This experiment helps to capture the features in a generalized sense of dysarthric classification. When the model is trained using UA-Speech corpus and tested using TORGO, there is a significant drop in performance. This is because the TORGO dataset consists of both words and sentences along with the different recording conditions. The experiments are performed without omitting the sentences as this experimentation pushes the generalization concept to its limits. From Table 5.3, the MGDCC outperforms the baseline features on both the experimental setup indicating that the feature set is able to capture dysarthric-based features in challenging conditions when the test data is completely different from the training dataset. Further, it can be observed from Table 5.3 that the performance of GDCC is poor than MFCC and LFCC, more so, than MGDCC; indicating that the modified GDF is indeed helping us to capture the formant structure better than GDCC.

Table 5.3: Cross-Database Evaluation using CNN Classifier.

Train	UA-Speech				TORGO			
Test	TORGO				UA-Speech			
	Accuracy	P	R	F1	Accuracy	P	R	F1
MFCC	33.85	34.08	35.22	35.14	42.03	45.22	42.05	43.57
LFCC	27.60	20.72	28.98	24.16	49.14	52.71	49.16	48.22
GDCC	28.31	31.72	23.61	27.07	29.58	29.13	33.04	30.97
MGDCC	**43.29**	**47.02**	**45.98**	**46.49**	**51.24**	**53**	**50**	**51.45**

5.4.6 Analysis of Latency Period

Latency period analysis is performed on MFCC, LFCC, and MGDCC feature sets on UA-Speech and TORGO corpora as shown in Fig 5.5. The latency period is estimated by computing the % classification accuracy w.r.t varying time duration of the speech segment. The latency period is varied from 50 frames (0.12 seconds) to 1200 frames (3 seconds). The latency period is the duration between the speech utterances produced to the system, and the response from the system in terms of % fold accuracy indicating the number of frames considered for the classification of an utterance. Hence, if the system performs well at lower latency, then it can be understood that the system classifies the utterance without needing a larger duration of the speech. It can be seen that the MGDCC features give a significant

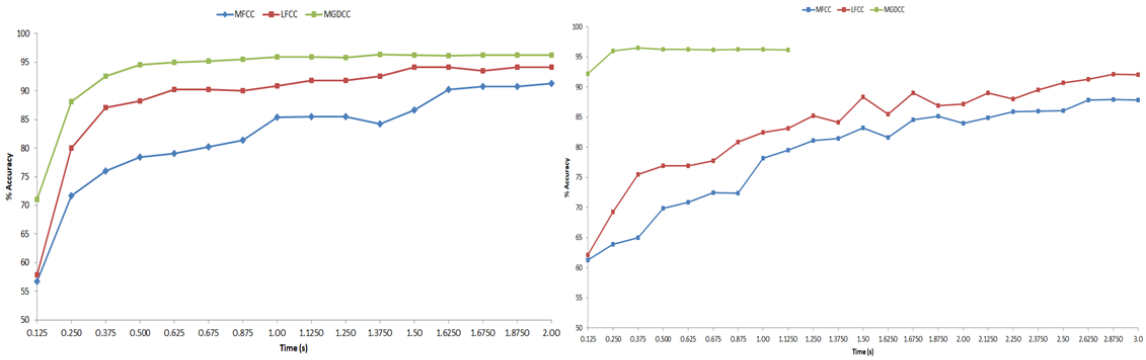

Figure 5.5: Latency Period Analysis on (a) UA-Speech Corpus, (b) TORGO Database, using CNN Classifier. After [69].

classification performance in the limited duration speech utterance of < 675 ms. On the contrary, both the MFCC and LFCC feature sets take a relatively longer duration to achieve comparable performance. A similar trend can be observed for the latency period using the TORGO database from Fig 5.5. The MGDCC features reached the maximum accuracy for the limited duration speech utterance of <250

ms.

5.4.7 Chapter Summary

This chapter discusses the application of phase-based MGDCC features for dysarthric speech. The parameter tuning of MGDCC is shown followed by the dimensionality tuning of the feature vector. Later, the proposed features are seen outperforming the magnitude-based features, which are considered the baseline features. Furthermore, the effect of the dynamic features for severity-level classification is observed. Finally, the chapter closes with the cross-database evaluation and latency period analysis. Since the 5-Fold score and test score lie in a similar range of values. The following results will represent the 5-Fold test accuracy alone. This also provides us insights into the average range of variation of the test accuracy. In the next chapter, we present the property of noise robustness of the modified group delay function. Later, the dysarthric speech detection task is performed using the proposed feature set.

CHAPTER 6
Noise Robustness of Modified Group Delay Function

6.1 Introduction

Numerous techniques have been proposed for the classification of dysarthria severity-level. However, these methods may exhibit degraded performance in the presence of background noise, resulting in inaccurate severity-level classification. To address this limitation, this chapter investigates the noise robustness of the group delay and modified group delay functions for dysarthria severity-level classification. The evaluation is conducted on the UA-Speech and TORGO corpora using a CNN classifier. Various stationary and non-stationary noises are employed to test the robustness of the proposed functions, including white and pink noise as stationary noises, and street and babble noises as non-stationary noises, at different Signal-to-Noise Ratio (SNR) levels.

Additionally, dysarthric speech detection is performed using Fourier transform phase-based features. The baseline features for comparison are magnitude-based Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). The experimental results demonstrate that the proposed features outperform the baseline features, even under severe signal degradation caused by noise. Moreover, the proposed feature set exhibits strong classification performance in distinguishing between speakers with very low severity-level and control speakers, surpassing the performance of the baseline features. This highlights the effectiveness of phase-based features.

The experiments involve an 80 % training and 20 % testing data split, and the model's performance is evaluated using the 5-Fold cross-validation technique. The results of the 5-Fold test accuracy are presented, demonstrating the robustness and efficacy of the proposed features in various noise conditions.

6.2 Additive Noise Robustness of Group Delay

In this sub-section, we analytically show the robustness of the group delay function to additive noise, which is also applicable to the modified group delay function [63]. Let $u(n)$ showcase a clean speech signal, degraded by adding additive noise, which is uncorrelated $v(n)$ with 0 mean, and σ^2 variance. Then, the noisy speech $z(n)$ can be expressed as:

$$z(n) = u(n) + v(n). \tag{6.1}$$

Taking the Fourier transform and obtaining the power spectrum we have,

$$P_z(\omega) = P_u(\omega) + P_v(\omega). \tag{6.2}$$

From eq.(6.2), there can be two mutually exclusive frequency regions, namely, *high SNR* and *low SNR*.

Low SNR Case

Considering a low SNR situation, i.e., $P_u(\omega) << \sigma^2(\omega)$ (where noise power is $\sigma^2(\omega)$ due to the assumption that noise is having 0 mean), we have:

$$P_z(\omega) = \sigma^2(\omega)(1 + \frac{P_u(\omega)}{\sigma^2(\omega)}). \tag{6.3}$$

Taking the logarithm on both sides and using the Taylor expansion, and ignoring the higher-order terms results in:

$$ln(P_z(\omega)) \approx ln(\sigma^2(\omega)) + \frac{P_u(\omega)}{\sigma^2(\omega)}. \tag{6.4}$$

Since $P_u(\omega)$ is a continuous and periodic function of ω, it can be expanded using the Fourier series. In particular,

$$ln(P_z(\omega)) \approx ln(\sigma^2(\omega)) + \frac{1}{\sigma^2(\omega)}\left[\frac{d_0}{2} + \sum_{k=1}^{+\infty} g_k cos(\frac{2\pi}{\omega_0}\omega k)\right], \tag{6.5}$$

Since $P_u(w)$ is a power spectrum, it is an even function, and the coefficient values of sine (basis) terms are zeros. To relate the spectral phase and magnitude with the cepstral coefficients, let us consider the Fourier transform representation of a

sequence b(n):

$$B(e^{j\omega}) = |B(e^{j\omega})|e^{j\theta(e^{j\omega})}. \tag{6.6}$$

Since the log-magnitude component is an even function, the resulting Fourier series expansion can be given by:

$$ln(|B(e^{j\omega})|) = \frac{c[0]}{2} + \sum_{n=1}^{+\infty} p[n]cos(\omega n). \tag{6.7}$$

From the properties of the Fourier phase spectrum, the phase spectrum is an odd function. Hence, the resulting Fourier series expansion is given by:

$$\theta(e^{j\omega}) = -\sum_{n=1}^{+\infty} p[n]sin(\omega n), \tag{6.8}$$

where p[n] is the n^{th} cepstral coefficient. Group delay coefficients are obtained by considering the negative logarithm of the unwrapped phase obtained in eq (6.8):

$$T(e^{jw}) = \sum_{n=1}^{+\infty} np[n]cos(\omega n). \tag{6.9}$$

From eqs (6.7) and (6.8), it can be observed that the phase and log-magnitude spectra of a signal are related through the cepstral coefficients. Assuming the additive noise as a minimum phase signal [31]. From eqs (6.7), (6.8), and (6.9), it can be observed that the group delay function can be extracted from the log-magnitude response by ignoring the DC term, and multiplying each coefficient by k. Applying this observation to eq.(6.5) we can obtain the group delay function as [63]:

$$T_z(\omega) \approx \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{+\infty} kg_k cos(\omega k). \tag{6.10}$$

Eq.(6.10) indicates that the group delay function is *inversely* proportional to the noise power in the regions with low SNR. This indicates that the group delay function preserves peaks and valleys well in the presence of additive noise and hence, helps in speech intelligibility.

High SNR Case

Now assume the case such that $P_u(\omega) >> \sigma^2(\omega)$, we have:

$$P_z(\omega) = P_u(\omega)(1 + \frac{\sigma^2(\omega)}{P_u(\omega)}). \tag{6.11}$$

Taking the logarithm on both the sides and using the Taylor series expansion results in:

$$ln(P_z(\omega)) \approx ln(P_u(\omega)) + \frac{\sigma^2(\omega)}{P_u(\omega)}. \tag{6.12}$$

Since $P_u(w)$ is a non-zero, continuous, and periodic function of ω, the same can be said about $\frac{1}{P_u(\omega)}$. Hence, both ln(.) and $(\frac{1}{\cdot})$ of eq.(6.12) can be expanded using the Fourier series, resulting in:

$$ln(P_z(\omega)) \approx \frac{d_0}{2} + \frac{\sigma^2(\omega)e_0}{2} + \sum_{k=1}^{+\infty}(g_k + \sigma^2(\omega)e_k)cos(\omega k), \tag{6.13}$$

where g_k's and $e_k's$ are the Fourier series coefficients of $ln(P_u(\omega)$ and $\frac{1}{P_u(\omega)}$, respectively.

Using eqs (6.7) and (6.8), we obtain the group delay function as:

$$T_z(\omega) \approx \sum_{k=1}^{+\infty} k(g_k + \sigma^2(\omega)e_k)cos(\omega k). \tag{6.14}$$

Eq.(6.14) indicates that the noise power $(\sigma^2(\omega))$ is negligible, when the signal power is higher than the noise power and the group delay function can be represented by only using the log-magnitude spectrum. Hence, from both the SNR cases, it can be concluded that the group delay spectrum follows the signal spectrum instead of the noise spectrum making it robust to the additive noise.

6.3 Spectrographic Analysis

Fig.6.2 represents the noisy speech signal corrupted using white noise at SNR level of -5 dB. From the time-domain plots, it is observed that the VOT region of the signal is completely masked by the addition of noise. From the group delay gram, the amount of noise at the lower frequencies is significantly less, when compared with magnitude-based features. This helps in the detection of the VOT due to the fact that the VOT is observed at lower frequencies. Additionally, the formant structure, and the resolution are preserved even in degraded conditions. Furthermore, MGDCC is found to boost signal energy along with the minimization of noise energy. This can be explained by the analytical analysis performed above, where it is seen that the modified group delay function follows the signal energy rather than the noise energy.

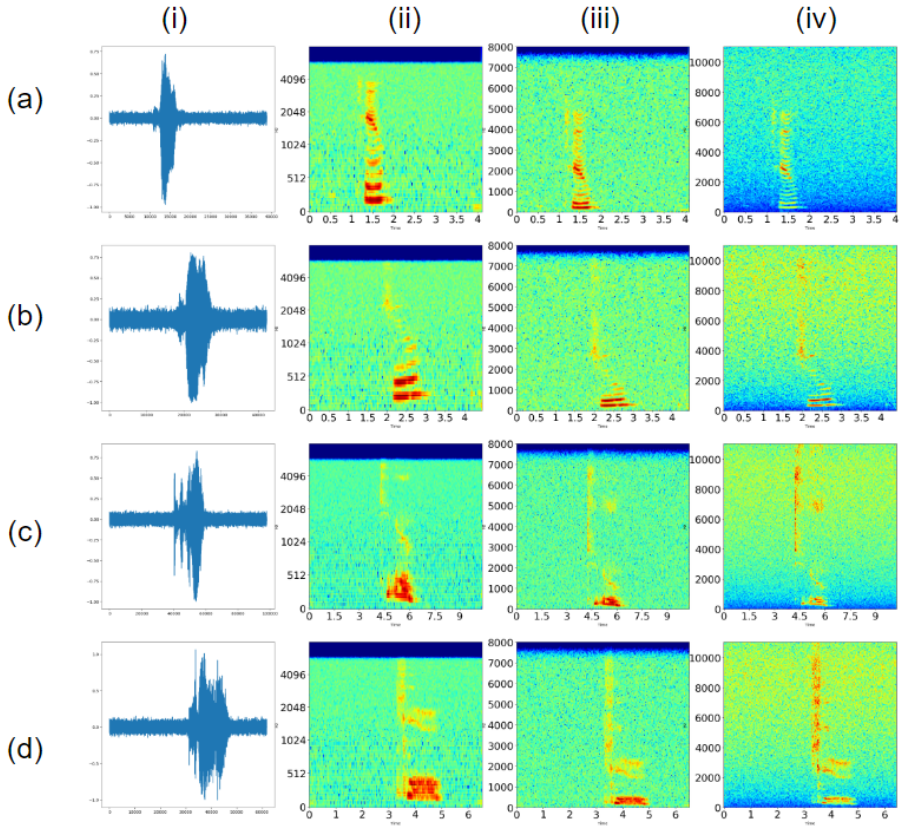

Figure 6.1: Fig. 5.2(i), Fig. 5.2(ii), Fig. 5.2(iii), and Fig. 5.2(iv) of Each Panel Depicts the Time-Domain Waveform, Mel Spectrogram, STFT Spectrogram, and Modified Group Delay-Gram of the Noisy Dysarthric Speech Signal of the Word *"to"*. Fig. 5.2(a), Fig. 5.2(b), Fig. 5.2(c), Fig. 5.2(d) Indicate Very Low, Low, Medium, and High Dysarthric Severity-Level Noisy Speech Signals.

6.4 Experimental Results

6.4.1 Results under Signal Degradation Conditions

The robustness of the proposed MGDCC feature set is tested using various noise types, such as white, pink, babble, and street noise with SNR levels of -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. Table 6.1 and Table 6.2 indicate the robustness of UA-Speech and TORGO corpora, respectively. When considered white noise for evaluation, due to the nature of AWGN, the noise gets added equally in all the frequency bands. From Fig 6.2, MGDCC outperforms MFCC and LFCC by a considerable margin. The MFCC performs poorly than the LFCC because of the amount of noise added in the higher frequency regions. The MFCC has fewer subband filters when compared with LFCC (due to Mel frequency wrapping) resulting in poor robustness for white noise. Considering that the signal

is degraded by the pink noise, which contains more noise power at lower frequencies, and less noise power at higher frequencies. The MGDCC feature set continues to show the noise robustness to the additive noise. MFCC performed better when compared with LFCC due to the nature of pink noise. Since the noise gets added in lower frequency regions, the higher number of filterbanks in MFCC helps it to become robust for the pink noise. Additionally, non-stationary noises such as street noise and babble noise are considered. The MGDCC feature continues to outperform the baseline features (MFCC and LFCC). These results indicate that the performance of the baseline features is degraded in the presence of stationary and non-stationary noise, whereas the performance of MGDCC remains intact across various noise types. These results prove the additive noise robust-

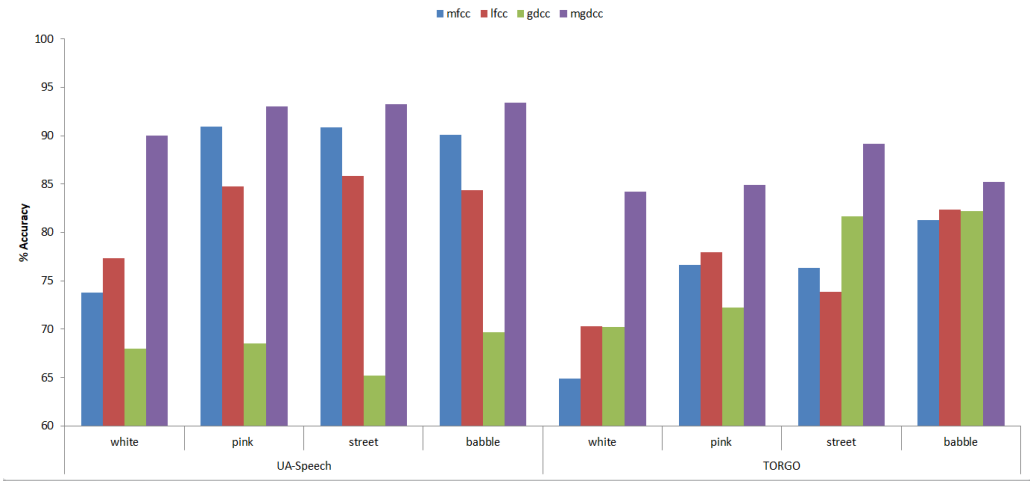

Figure 6.2: Accuracy at Low SNR Levels using CNN Classifier.

ness property, and also that the group delay spectrum is known to emphasize the *signal* spectrum rather than the *noise* spectrum. It can also be explained by the fact that the MGDCC feature set pushes the zeros into the unit circle (in the Z-plane) in an attempt of making the signal a minimum phase, which may also help in the suppression of noise. The fact that the group delay function gives a larger peak at formant frequencies, when compared with magnitude-based features might also be a reason for better performance under noisy conditions.

6.4.2 Results under Severe Degradation Conditions

The experiments of severe signal degradation are performed using white noise, and a CNN classifier on the UA-Speech corpus. From Table 6.3, as the SNR level is dropped to as low as -40 dB, the MFCC and LFCC clearly struggle with the accuracy dropping as low as 39. 78 % and 24 %, respectively. On the contrary,

Table 6.1: % Accuracy for Various Noise Types Across Various SNR Levels using CNN Classifier on UA-Speech Corpus.

Noise Type	Features	SNR (dB)						
		-10	**-5**	**0**	**5**	**10**	**15**	**20**
White	**MFCC**	66	67.31	81	80.84	81.38	92.68	94.48
	LFCC	**61.30**	66.29	89.68	92	92.86	93.71	93.46
	GDCC	**68.19**	68.40	68.51	69.14	69.50	69.85	70.29
	MGDCC	**86.24**	**90.28**	**91.55**	**92**	**94**	**94.48**	**94.80**
Pink	**MFCC**	89.82	90.56	91.40	92	94.52	94.52	95.22
	LFCC	**75.54**	77.70	92.47	93.18	94.16	94.60	95.54
	GDCC	**67.71**	68.12	69.76	65	64.31	69.90	70.11
	MGDCC	**91.06**	**92.40**	**93.32**	**95.19**	**95.26**	**95.41**	**95.58**
Street	**MFCC**	87.06	91	91.80	93.49	94	94.24	94.24
	LFCC	**80.60**	80.38	89.85	92.40	93.10	93.47	94.84
	GDCC	**63.92**	94.92	67.31	68.26	61.05	68.14	68.51
	MGDCC	**93.49**	**94.38**	**94.38**	**94.82**	**95.30**	**95.73**	**95.79**
Babble	**MFCC**	88.16	90.58	91	90.70	91.90	91.20	91.30
	LFCC	**79.85**	80.71	88.47	93.49	93.78	95.37	95.97
	GDCC	**67.71**	69.46	69.76	66.60	69.99	70.84	70.88
	MGDCC	**92.07**	**93**	**94.38**	**94.90**	**95.01**	**95.08**	**95.68**

Table 6.2: % Accuracy for Various Noise Types Across Various SNR Levels using CNN Classifier on TORGO Corpus.

Noise Type	Features	SNR (dB)						
		-10	**-5**	**0**	**5**	**10**	**15**	**20**
White	**MFCC**	62.01	64.98	67.76	70.47	72.11	75.42	82.01
	LFCC	70.22	70.59	70.13	71.30	71.54	71.75	74.71
	GDCC	64.98	71.02	71.70	74.20	74.95	77.66	80.43
	MGDCC	**80.56**	**84.21**	**87.69**	**87.79**	**87.87**	**88.95**	**91.79**
Pink	**MFCC**	74.63	77.35	75.70	80.34	80.44	81.51	84.72
	LFCC	77.35	78	78.54	81.77	83.54	90	91.14
	GDCC	71.60	74.51	76.02	73.88	80.60	79.93	82.15
	MGDCC	**80.63**	**86.11**	**86.67**	**87.63**	**88.39**	**89.02**	**90.77**
Street	**MFCC**	76.65	76.89	75.45	75.45	76.08	81.24	83.53
	LFCC	74	74.91	72.61	79.05	79.49	81.51	86.23
	GDCC	80.96	82.11	82.01	85	87.57	84.41	85.14
	MGDCC	**88.13**	**89.33**	**90.12**	**90.15**	**90.36**	**92.55**	**90.11**
Babble	**MFCC**	77.15	82.14	84.54	83.92	84.72	86.11	88.10
	LFCC	81.19	83.09	82.9	80	83.49	85.92	87.79
	GDCC	80.85	84.88	83.91	86	91.89	88.95	89.58
	MGDCC	**87.12**	**84.52**	**84.03**	**86.18**	**86.75**	**90.22**	**90.66**

Table 6.3: % Accuracy for Signal Degraded using White Noise at Severe SNR Levels using CNN Classifier on UA-Speech Corpus

SNR (dB)	**MFCC**	**LFCC**	**MGDCC**
-15	59.97	57.74	**83.73**
-20	50	44.9	**77.50**
-25	41.81	33.15	**68.19**
-30	39.92	31.24	**63.07**
-35	39	26.62	**52.29**
-40	39.78	24	**52**

MGDCC outperforms MFCC and LFCC by **12.22** % and **28** %, respectively. This experiment shows the extent of the robustness of the proposed feature set.

6.4.3 Dysarthric Speech Detection

From Fig 6.3 and Fig 6.4, MGDCC shows significantly higher accuracies in the classification of normal speech w.r.t very low and low severity-levels. This result encourages the use of MGDCC features due to the fact that the majority of the current dysarthria systems benefit from the detection of higher severity-levels, which do not need much information. Hence, it can be inferred that the MGDCC captures the distinguishing features between normal and dysarthria speech well even for lower severity-levels. On the contrary, GDCC performs poorly for low

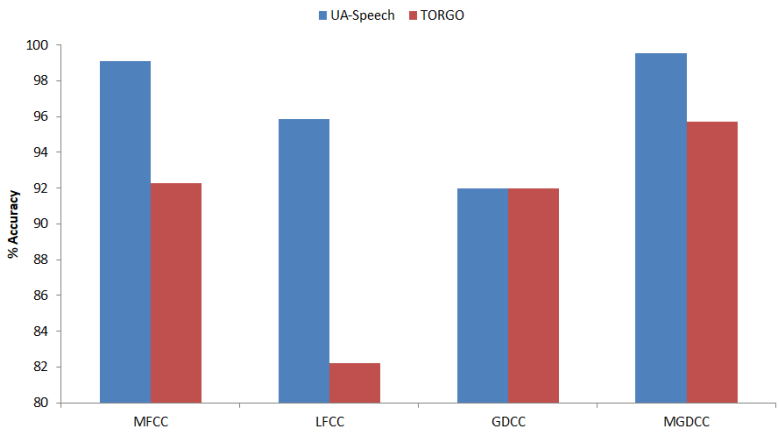

Figure 6.3: Dysarthric Speech Detection on Both Datasets using CNN Classifier.

severity-levels due to the fact that the group delay function fails to capture formant information due to the unwanted peaks in the spectrum.

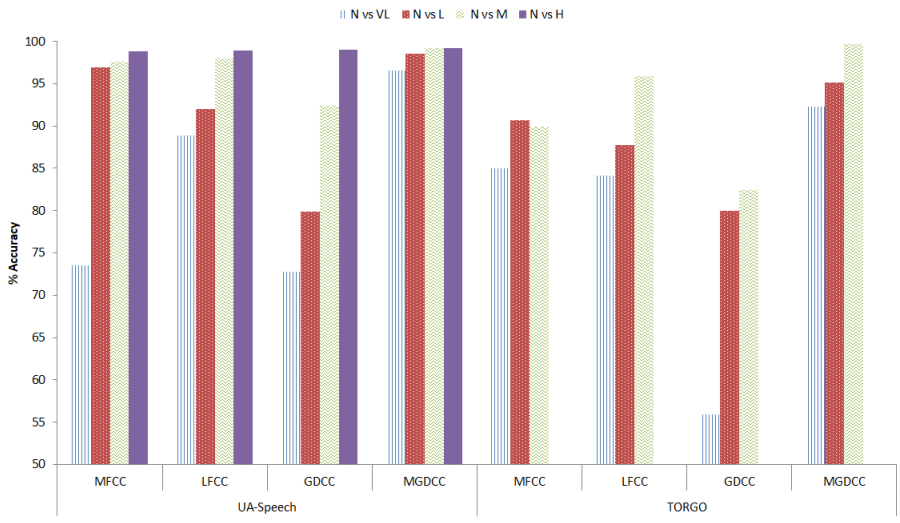

Figure 6.4: Dysarthric Speech Detection on Both Datasets using CNN Classifier.

6.4.4 Chapter Summary

This chapter demonstrated the noise robustness of modified group delay-based cepstral features (MGDCC). The analytical explanation is supported by the results obtained using experiments performed by considering stationary and non-stationary noises at various SNR levels for severity-level classification. Further, dysarthric speech detection is performed. The next chapter discusses the application of phase-based features for speech emotion recognition task.

CHAPTER 7
Modified Group Delay Features for Speech Emotion Recognition

7.1 Introduction

As technological advancements progress, dependence on machines is inevitable. Therefore, to facilitate effective interaction between humans and machines, it has become crucial to develop proficient techniques for Speech Emotion Recognition (SER). This chapter proposes the feature set, namely, MGDCC for the SER task. This chapter shows the ability of phase-based features to classify emotions. The feature sets are applied to the German language-based EMO-DB database, and the results are obtained on the CNN classifier. The magnitude-based Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC) are baseline features for this study. From the results, it can be observed that the proposed features surpass the baseline features comfortably. Furthermore, the noise robustness of the MGDCC feature set is also explored with stationary and non-stationary noise types. Additionally, to check the practicality of the proposed feature set, latency period analysis is performed. All the mentioned features are evaluated by keeping a window size of 25 ms and a hop length of 10 ms, F_{min} = 100 Hz, and octave resolution of 14. This work uses the leave one speaker out technique in order to check the speaker independence of the feature set. The training data consists of speech samples from 9 speakers, while the test data contains 1 speaker. From the results, it can be noticed that the phase-based features outperform the baseline features. The analytical explanation of the robustness of additive noise is practically proven through experimentation.

7.2 Motivation of Phase-Based Features for Emotion Recognition

Phase-based features capture temporal information and variations of a speech signal which are important factors in SER. Some of the vital features for emotion recognition are prosody, timing and the rhythm of speech, and non-verbal cues, such as breathiness, which are captured by the phase-based features. Fig 7.1 represents the Mel-spectrogram, spectrogram, and group-delay-gram of male and female speakers for anger, happy, sad, and neutral emotions. The formant reso-

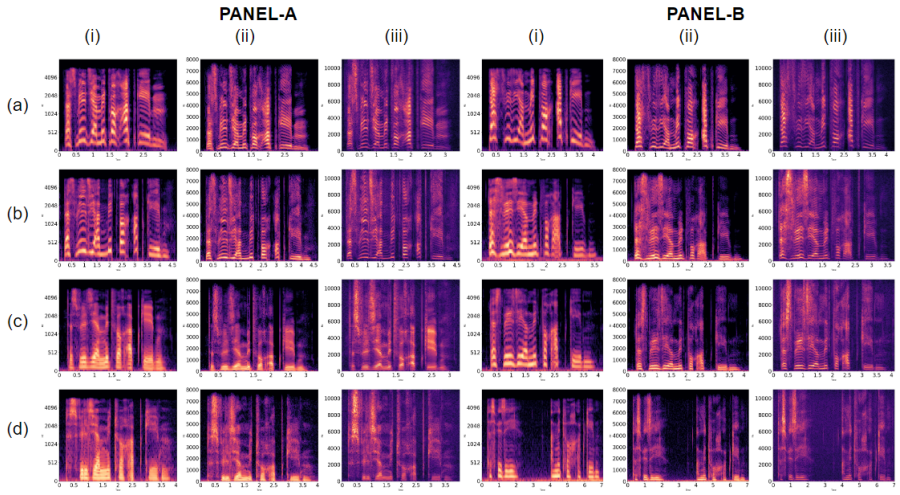

Figure 7.1: Panel-A and Panel-B Represent Plots for Male and Female Speakers, Respectively. (i), (ii), and (iii) Represents Mel spectrogram, Spectrogram, and Group-delayGram. (a), (b), and (c) Represents Anger, Happy, Sad, and Neutral Emotions, Respectively.

lution in Mel-spectrogram at lower frequencies is good but as we move towards higher frequencies, we can see the resolution getting poorer. It can be observed from the plots that the fine structure of the formants that can be observed in the magnitude spectrum (Panel-A) can also be seen in the spectrogram obtained by the modified group delay spectrum. Hence, there is no information loss, while using phase-based cepstral coefficients. This is due to the fact that the denominator term at the formant frequencies becomes 0 (as the pole radius approaches the unit circle in the Z-plane) resulting in peaks that give higher-resolution formants. Additionally, phase features are able to capture irregularities in the speech signal. The presence of turbulence in a speech signal changes with emotion and these irregularities are captured better through phase features rather than the magnitude spectrum.

7.3 Experimental Result

7.3.1 Parameter Tuning of Modified Group Delay Function

MGDCC involves two constraint parameters, alpha (α), and gamma (γ). The parameters are fine-tuned using a greedy search algorithm i.e., the performance is optimized wrt the performance of SER. The parameters are varied from 0 to 1 with a step size of 0.1. The evaluation is done based on the test scores of the CNN classifier. The best test accuracy is achieved at parameter values of α **= 0.1** and γ **= 0.1**, resulting in an accuracy of **79.49** %. The effect of parameters can be seen in Fig 7.2.

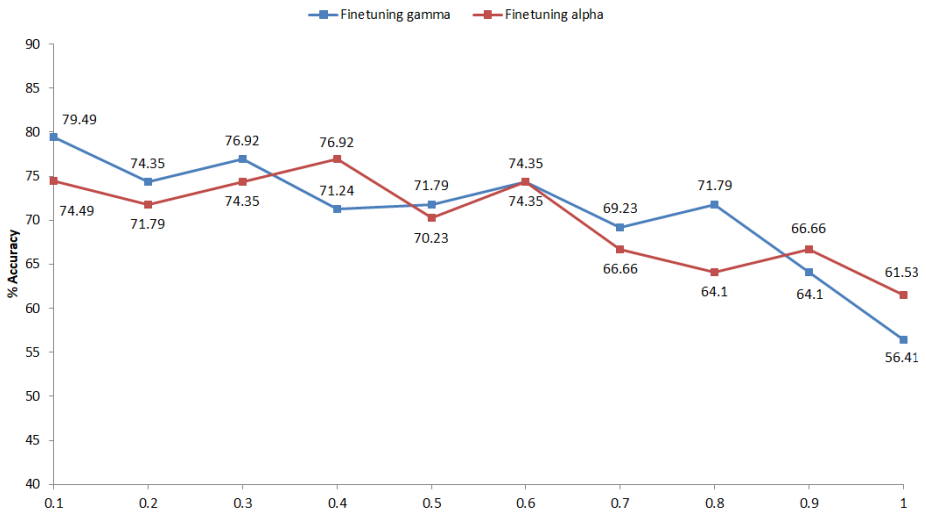

Figure 7.2: Tuning Parameters α and γ using Greedy Search Technique for Emotion Recognition.

7.3.2 Results for Emotion Recognition

The MGDCC feature outperforms the magnitude-based features, MFCC and LFCC by a margin of **7.7** % and **5.14** %, respectively. The reason behind LFCC outperforming MFCC might be because of the importance of higher frequency information for certain emotions such as anger and happy. These emotions contain a higher pitch leading to formants occurring at a higher frequency. The resolution of the LFCC feature is better than the MFCC features at higher frequencies. This might be because of the high-resolution property of the modified group delay function. The MGDCC outperforms all the features due to its ability to capture the formants with high resolution at low and higher frequencies. However, GDCC

fails to achieve similar performance. This is because of the noisy structure resulting from the GDCC occurring from the presence of zeros close to or outside the unit circle in the Z-plane. These spikes cause formant masking and also hamper speech intelligibility, thereby making it difficult to obtain valuable features for the classification task.

Table 7.1: Accuracy of EMO-DB Dataset using CNN Classifier.

Features	**Test Acc.**
MFCC	71.19
LFCC	74.35
GDCC	56.41
MGDCC	**79.49**

7.3.3 Results under Signal Degradation Conditions

The robustness of the proposed features is tested using various noise types, such as white, pink, babble, and street noise with SNR levels of -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. When we consider additive white noise for evaluation, due to the nature of AWGN, the noise is distributed across all the bands of frequency. From Table 2, at the low SNR levels, MGDCC clearly outperforms both the magnitude-based features, MFCC and LFCC by a significant margin of **3.41 %**, **10.25 %**, respectively. Similarly, at higher SNR values, MGDCC outperforms baseline features MFCC and LFCC by **17.95 %**, **7.79 %**, respectively. Considering that the signal is degraded by the pink noise, which has higher noise power in lower frequencies rather than the higher frequencies, the MGDCC feature set outperforms both MFCC and LFCC features. Additionally, when considered non-stationary noises (noises which vary w.r.t time), such as street noise or traffic noise, and babble noise are considered. The MGDCC noise robustness is observed in any kind of noise. These results indicate that the performance of the baseline features is degraded in the presence of stationary and non-stationary noise, whereas the performance of MGDCC remains intact across various noise types.

These results illustrate the additive noise robustness property, and also that the group delay spectrum is known to emphasize the signal spectrum rather than the noise spectrum. It can also be explained by the fact that the MGDCC feature set pushes the zeros into the unit circle (in the Z-plane) in an attempt of making the signal a minimum phase, which may also help in the suppression of noise. Additionally, it can be noted that LFCC and MFCC are not equally robust in white

Table 7.2: % Accuracy for Various Noise Types Across Various SNR Values using CNN Classifier on EMO-DB Dataset.

Noise Type	Feature Set	SNR Level(dB)					
		-10	**-5**	**0**	**5**	**10**	**15**
White	**MFCC**	69.23	76.92	74.35	43.58	82.05	43.58
	LFCC	71.79	64.10	64.10	58.97	71.79	69.23
	MGDCC	**76.92**	**79.48**	**74.35**	**71.79**	**76.92**	**74.35**
Pink	**MFCC**	41.79	38.46	41.02	43.58	70.35	71.79
	LFCC	71.79	66.66	66.66	61.53	71.79	71.79
	MGDCC	**74.35**	**69.23**	**71.79**	**71.79**	**71.79**	**74.35**
Street	**MFCC**	74.35	76	76.92	79.48	88.48	82.05
	LFCC	74.35	70.23	79.48	71.79	76.92	79.48
	MGDCC	**75.53**	**80**	**81.66**	**81.66**	**71.79**	**86.66**
Babble	**MFCC**	74.35	76	79.48	79.48	79.48	79.48
	LFCC	61.53	66.66	79.48	76.92	79.48	79.48
	MGDCC	**79.48**	**81.29**	**82.05**	**81.66**	**81.66**	**82.66**

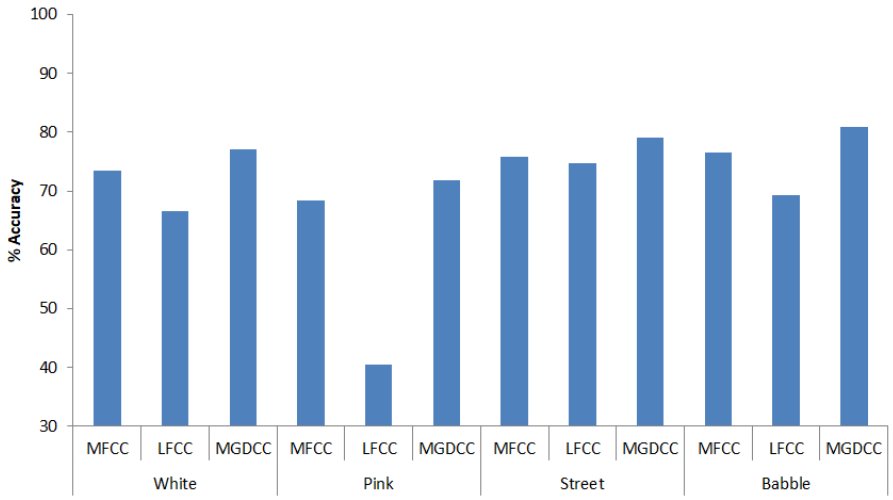

Figure 7.3: Performance of Features at Low SNR Values using CNN Classifier.

noise as the energy in higher frequency speech regions is weak making it more susceptible to noise corruption. The LFCC contains more subband filters at higher frequencies than MFCC, making it less robust to white noise. As the noise power decreases, the LFCC feature set still outperforms MFCC due to its linearly-spaced subband filters instead of the Mel filterbank. This reasoning also explains the comparable performance of MFCC to LFCC, when the signal is corrupted with pink noise.

7.3.4 Analysis of Latency Period

In this thesis, we investigated the latency period of the MGDCC feature set in comparison to the baseline features, i.e., MFCC and LFCC. To evaluate the performance of CNN based on different feature sets, we measured the accuracy % wrt the latency period, as depicted in Fig.7.4. The latency period denotes the time

elapsed between the utterance of speech and the system's response, expressed as a percentage fold accuracy that represents the number of frames utilized for utterance classification. Therefore, if the system demonstrates superior performance at

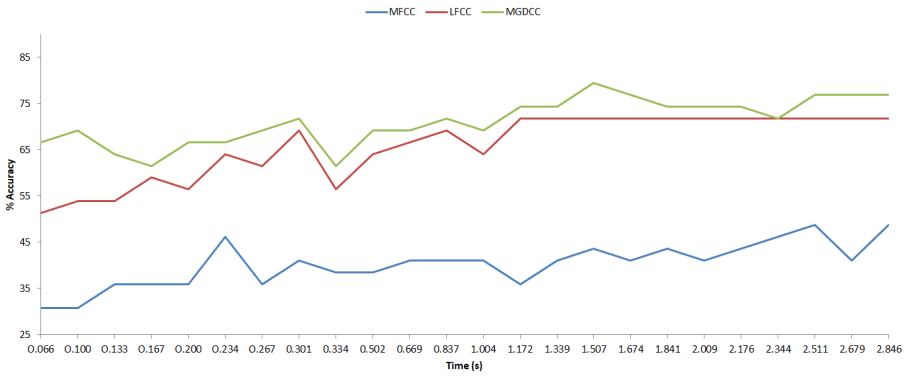

Figure 7.4: Analysis of Latency Period for Various Feature Sets using CNN Classifier.

lower latency periods, it implies that it can classify the speech utterance effectively without requiring a prolonged duration of speech. The duration of utterance is up to *3* sec and is plotted at an interval of *0.5* sec. It is observed that MGDCC features give significant classification performance throughout, the highest accuracy being *79.48* % at 1.5 sec. On the contrary, the baseline features constantly down-perform and for MFCC, it takes a longer duration to achieve comparable performance. This encourages the practical suitability of the proposed MGDCC feature set.

7.3.5 Chapter Summary

In this study, phase-based vocal tract system features were proposed for the SER task. Other state-of-the-art spectral features MFCC and LFCC were used for comparison. The objective was to capture the irregularities in the speech signal and the formant structure better for efficient SER. MGDCC also proved to perform well for stationary and non-stationary noise-added datasets due to its additive noise robustness property. The significance of linear filterbanks over Mel filterbanks was observed for SER. The practical suitability of MGDCC was also calculated and promising results were seen. The next chapter showcases the application of MGDCC features for voice liveness detection (VLD) through pop noise detection.

CHAPTER 8
Modified Group Delay Features For POP Noise Detection

8.1 Introduction

Automatic Speaker Verification (ASV) systems play a crucial role in security systems. However, these systems are susceptible to various spoofing attacks, including the playback of pre-recorded or synthetic speech. To mitigate such vulnerabilities, this paper explores the utilization of pop noise analysis as a feature in voice-liveness detection algorithms. Pop noises, short-duration acoustic disturbances, often occur during the production or recording of live speech but are absent or different in pre-recorded or synthetic speech. Leveraging the unique temporal and spectral characteristics of pop noises, we propose incorporating them as discriminative cues for voice liveness detection. By analyzing the occurrence, duration, and timing of pop noises within the speech signal, the system can differentiate live speech from spoofed or synthetic speech. In this study, we propose the detection of pop noise using the phase-based feature, Modified Group Delay Cepstral Coefficients (MGDCC). The key idea behind employing MGDCC features is that pop noise being breath sounds, create irregularities in the speech signal and use to capture these irregularities via phase-based MGDCC features. The proposed feature is compared against the baseline features, namely, Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC). All the mentioned features are evaluated by keeping a window size of 25 ms and a hop length of 10 ms, F_{min} = 40 Hz, and octave resolution of 14. The experiments are performed on the training data of the POCO dataset using CNN as a classifier. The results are evaluated using 5 -Fold CV accuracy scores. Furthermore, the practicality of the proposed feature set is checked using latency period analysis. The results indicate that the proposed features outperform both the baseline features by a large margin indicating its ability to capture the energy in lower frequencies.

8.2 Motivation of Phase-Based Features for VLD

Row (a) of Fig 8.3 represents genuine speech that contains pop noise and (b) represents spoofed speech or recorded speech which does not contain pop noise for the word "chip". From the time domain plot, we can see the pop noise occurring at the end of the utterance of the word. In general, the pop noise occurs at lower frequencies as same as the VOT energy seen in previous chapters.

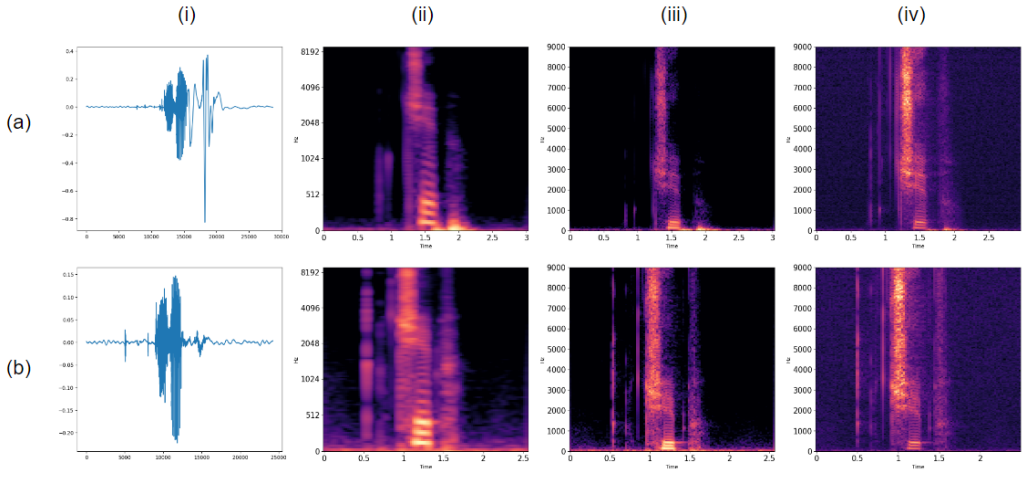

Figure 8.1: (a) and (b) Represents Plots for Genuine and Spoof Speech for the Word *"chip"*. (I), (ii), (iii), and (iv) Shows the Time-Domain, Mel-Spectrogram, STFT-Spectrogram, and Group-Delaygram, Respectively.

Since MGDCC was able to capture the information of VOT for control and dysarthric speakers, it motivated to use the phase-based MGDCC features for the pop noise detection task. One main differentiating feature across MEL-spectrogram, spectrogram, and group-delaygram is the presence of additional energy at lower frequencies in both genuine and spoofed speech which is absent in the group-delaygram. This might be crucial because the presence of additional energy at lower frequencies might mask the pop noise information which is not possible in the MGDCC features.

8.3 Experimental Results

8.3.1 Parameter Tuning of Modified Group Delay Function

MGDCC involves two constraint parameters, alpha (α), and gamma (γ). The parameters are fine-tuned using a greedy search algorithm i.e., the parameters are optimized wrt the performance of the VLD system. The parameters are varied

from 0 to 1 with a step size of 0.1. The evaluation is done based on the fold CV scores using CNN classifier. The best fold accuracy of **88.28** % is achieved for the parameters α = 0.22 and γ = 0.09, respectively.

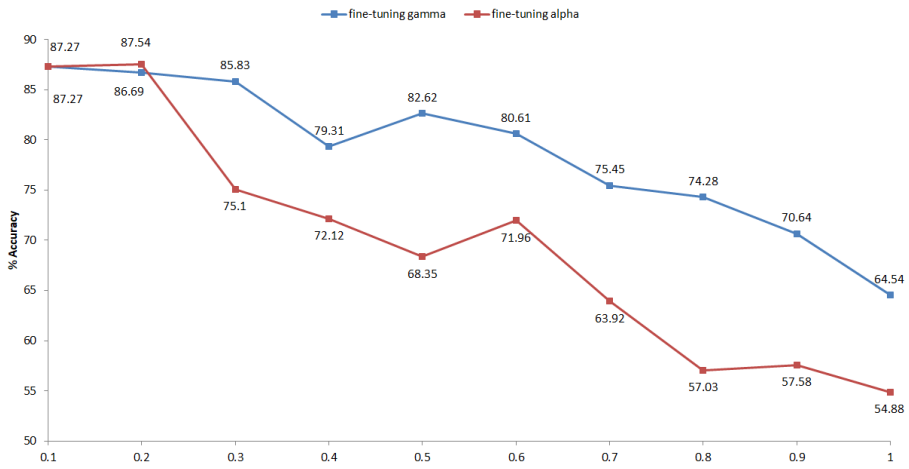

Figure 8.2: Parameter Tuning of α and γ Using Greedy Search Algorithm for Pop-Noise Detection.

8.3.2 Results for POP Noise Detection

From Table 8.1, it can be observed that the MGDCC outperforms the baseline magnitude-based features (MFCC and LFCC) by a margin of **36.07** % and **20.63** %, respectively. Hence, the phase-based cepstral features capture crucial information about the formant structure that aids in the classification. Among the baseline features, LFCC outperforms MFCC, which acts as a solid proof for the spectrographic analysis. Since the spectrograms are resulting in unwanted low frequency energies, which mask the pop noise information, MFCC performance drops heavily due to its good low-frequency resolution making the noisy energy much larger. On the other hand, traditional group delay function features do *not*

Table 8.1: % Accuracy of MGDCC on POCO Dataset using CNN Classifier.

Features	**Test Acc.**
MFCC	52.51
LFCC	67.64
GDCC	60.75
MGDCC	88.28

fare well. This is due to large spikes produced by the mixed-phase signals, such as speech. The presence of significant spikes in the group delay spectrum obscures

the formant information, thereby rendering these features less informative for the pattern classifier.

8.4 Analysis of Latency Period

In this chapter, we investigated the latency period of the MGDCC feature set in comparison to the baseline features, i.e., MFCC and LFCC. To evaluate the performance of CNN based on different feature sets, we measured the accuracy % wrt the latency period, as depicted in Fig.7.4. The latency period denotes the time elapsed between the utterance of speech and the system's response, expressed as a percentage fold accuracy that represents the number of frames utilized for utterance classification. Therefore, if the system demonstrates superior performance at

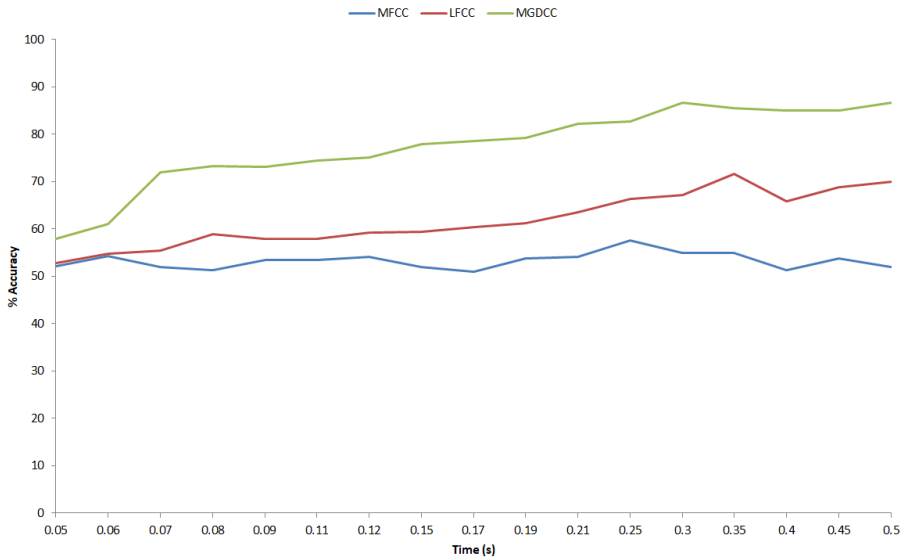

Figure 8.3: Latency Period Analysis of Pop Noise Detection using CNN Classifier.

lower latency periods, it implies that it can classify the speech utterance effectively without requiring a prolonged duration of speech. From the plot, it can be observed that the MGDCC achieves the maximum accuracy with 120 frames or *300 ms* speech sample. The MFCC and LFCC fail to achieve a saturation point. This experiment indicates the practicality of the proposed feature set.

8.5 Chapter Summary

This chapter discusses the application of MGDCC for pop noise detection, which is used for Voice Liveness Detection (VLD) tasks. The experiments are performed

on a subset of the POCO dataset and are evaluated using CNN classifier and a 5-fold cross-validation accuracy as the metric. The experiments proved that the phase-based features capture the formant information with a better resolution, when compared with magnitude-based (MFCC and LFCC) features. To test the practicality of the proposed feature set, the latency period analysis is performed. The next chapter proposes a novel technique of time-averaging features for infant cry classification task.

CHAPTER 9
Time-Averaged Features for Infant Cry Classification

9.1 Introduction

This chapter explores the effect of time-averaged features on the infant cry classification. The infant cry classification is a very relevant problem. The only way an infant can communicate is *cry*. The feature vectors Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) derived from STFT are used in this work. In this work, time-averaging is applied to the feature sets at various window sizes and is evaluated using machine learning classifiers which are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). The experiments are performed using Repeat Fold Cross Validation on the *Baby Chillanto* dataset using various features vectors. This method managed to achieve a high accuracy using simple machine learning classifiers.

9.2 Motivation

The cry signal considered in the dataset is samples of 1 second long. Considering the limited examples of the dataset and computational power required for the deep learning classifiers, the time averaging method is proposed. The proposed method achieves higher accuracy while using simple machine learning classifiers. The variations across the time-axis for a cry are minimal, when compared with adult speech, which contains many acoustic features while spelling out a word. The only changes in a baby cry would be the breathing pattern and the pitch, which are captured across the spectral-axis. Therefore, the features undergo time axis averaging, resulting in a reduction to 1-D. Figure 9.1, specifically Panel-I and Panel-II, presents the features generated using Librosa [47] for normal and pathological infant cries. This representation effectively captures the log magnitude

spectrum as Fourier transforms of cepstrum [18]. In Figure 9.1(a), static MFCC representations are depicted, while Figure 9.1(b) showcases dynamic MFCC representations. Additionally, Figure 9.1(c) illustrates the LFCC representations, and Figure 9.1(d) represents the cepstral coefficient representations.

Observing Figure 9.1(a) and Figure 9.1(b), it becomes evident that there are discernible differences in F0 and its harmonics among the classes of infant cry signals. The noticeable differences are also seen in LFCC features, as shown in Figure 2(c). However, the disparities are more pronounced in the dynamic MFCC representations compared to the static MFCC and LFCC representations. This observation can be attributed to the dynamic MFCC's ability to provide discriminative features across the entire frequency band. Furthermore, the results obtained through 10-fold cross-validation support the notion that dynamic MFCC yields the highest classification accuracy in this study. However, it should be noted that dynamic MFCC also contains redundant information when compared to static MFCC. As a result, the average accuracy of static MFCC exceeds that of dynamic MFCC. On the other hand, the features captured by cepstral coefficients (CC) lack sufficient discriminative power, making it challenging for classifiers to classify using these features.

9.3 Experimental Results

9.3.1 Parameter Tuning of Machine Learning Models

The parameters are tuned using a grid search algorithm with an accuracy metric. Table 9.1 shows the best parameters for all the machine learning classifiers used in this chapter.

Table 9.1: Parameter Tuning of Classifiers. After [67].

Classifier	Parameters	Static MFCC	Dynamic MFCC	LFCC	CC
KNN	**# Neighbors**	3	3	3	3
SVM	**C**	0.1	1	10	100
RF	**Max_depth**	20	50	10	50
	Sample_leaf	1	1	1	1
	Estimators	300	150	300	150

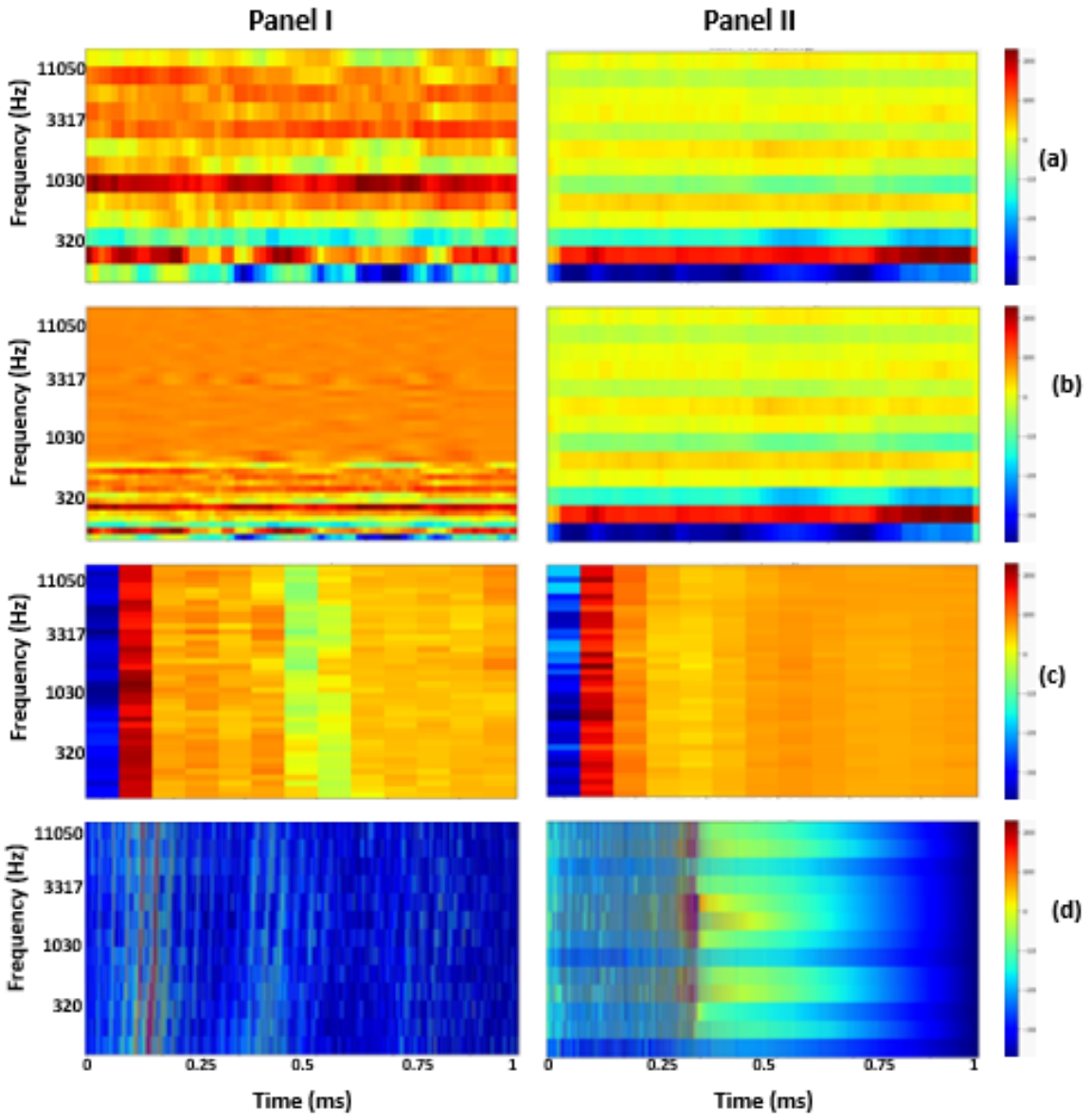

Figure 9.1: The Normal and Pathology Infant Cry Analysis are Shown in Panel-I and Panel-II. Fig. 9.1(a) Shows the Static MFCC Features, Fig. 9.1(b) Shows the Dynamic MFCCs, Fig. 9.1(c) shows the LFCC Features, and Fig. 9.1(d) Represents the Cepstral Coefficient Representations. After [67].

9.3.2 Results for Infant Cry Classification

The results are evaluated using Repeat Fold CV Technique. The static and dynamic coefficients of MFCC performed similarly resulting in an average accuracy across all the classifiers of 95.22 % and 93.71 %. The dynamic MFCC features can act as redundant-based features indicating it can decrease the performance of some classifiers, like SVM with a soft margin [12]. Overall, the dynamic MFCC capture the temporal trajectory of MFCC by incorporating velocity and acceler-

ation coefficients. Consequently, the inclusion of these dynamic MFCC features in the feature vector can introduce redundancies that may negatively impact the performance of classifiers.

The mean accuracy of LFCCs averaged over folds is 94.17 %. LFCCs utilize a linear filterbank, which results in better resolution at higher frequencies compared to the logarithmic resolution of MFCCs at higher frequencies. The average accuracy of the LFCC feature vector across all classifiers is higher than that of dynamic MFCCs but lower than that of static MFCCs. This suggests that the higher frequencies contain significant information and cannot be disregarded. The average accuracy of MFCC indicates that the lower frequency information which generally consists of the breathing information is also equally important. The results also highlight the impact of redundancy on the accuracy of dynamic features when compared to static MFCCs and LFCCs. The cepstral coefficient feature vector exhibits the lowest classification accuracy since it lacks the filterbanks present in MFCCs and LFCCs feature sets.

When it comes to the classifiers, each of them handles redundant data differently, as their classification techniques vary. The SVM classifier utilizes a linear kernel's decision boundary to classify between classes. In contrast, the KNN classifier applies a clustering concept by assigning a label based on the majority vote of its neighbours. The RF classifier assigns a label based on the majority vote from all the decision trees' outputs. Consequently, the performance of certain classifiers decreases when transitioning from a static MFCC to a dynamic MFCC feature vector, while others remain unaffected.

The performance of the SVM classifier using a linear kernel depends greatly on the effectiveness of feature extraction. In cases where feature extraction is not executed properly, the classification results tend to be unsatisfactory. This classifier is notably sensitive to redundant data. In contrast, the KNN classifier heavily relies on accurately extracted features and employs a clustering technique for classification. Therefore, when feature extraction is carried out correctly, KNN surpasses the linear SVM in performance. This is due to KNN's clustering-based classification, which renders the linear decision boundary ineffective, as evidenced by the results.

When comparing classifiers, the Random Forest (RF) classifier strives to outperform both the SVM and KNN classifiers across various feature extraction techniques. It achieves higher classification accuracy across all feature vectors by integrating multiple uncorrelated decision trees. As the number of decision trees increases, the RF classifier's ability to make accurate predictions also improves.

However, handling redundant data poses a challenge for the RF classifier, as the importance score can potentially lead the model astray.

The mean classification accuracy of the KNN classifier across all feature extraction techniques is 89.42 %. The optimal number of neighbours for the KNN classifier, determined through grid search, is 3, resulting in the best performance. On the other hand, the average classification accuracy of the Random Forest (RF) classifier across all feature extraction techniques is 90.29 %. The parameters, including maximum depth, minimum sample leaf, and the number of estimators, are fine-tuned using a grid search algorithm. They are set to 20, 1, and 300, respectively, achieving the best overall accuracy of 98.27 %.

In contrast, the SVM classifier with a linear kernel demonstrates an average classification accuracy of 78.82

By examining Table 9.2 and Table 9.3, the impact of window size (20ms and 55ms) on the accuracy. When the window size is increased, the temporal resolution decreases while the resolution in the frequency domain increases. Therefore, as we raise the window size from 20 ms to 55 ms, we notice an improvement in accuracy across the classifiers. This suggests that temporal information can be deemed less crucial in comparison.

Table 9.2: Accuracy for featured with a window size of 20 ms. Average Accuracy across Rows Indicate the Classifier Accuracy and Across Column Indicate Feature Accuracy. After [67].

Model	**Static MFCC**	**Dynamic MFCC**	**LFCC**	**CC**	**Average acc.**
KNN	97.98	98.38	96.74	61.28	88.59
RF	97.66	96.49	96.78	67.98	89.72
SVM linear	86.91	86.99	87.67	58.30	79.96
Average acc.	94.18	93.95	93.73	62.52	

Table 9.3: Accuracy for features with a window size of 55 ms. Average Accuracy across Rows Indicate the Classifier Accuracy and Across Column Indicate Feature Accuracy. After [67].

Model	**Static MFCC**	**Dynamic MFCC**	**LFCC**	**CC**	**Average acc.**
KNN	98.42	**98.48**	97.50	63.30	89.42
RF	97.92	96.61	97.27	69.37	90.29
SVM linear	85.44	86.07	87.75	56.05	78.82
Average acc.	93.92	93.72	94.17	62.90	

Due to the relatively low significance of temporal information in distinguishing between normal and pathological cries, an approach was employed where the

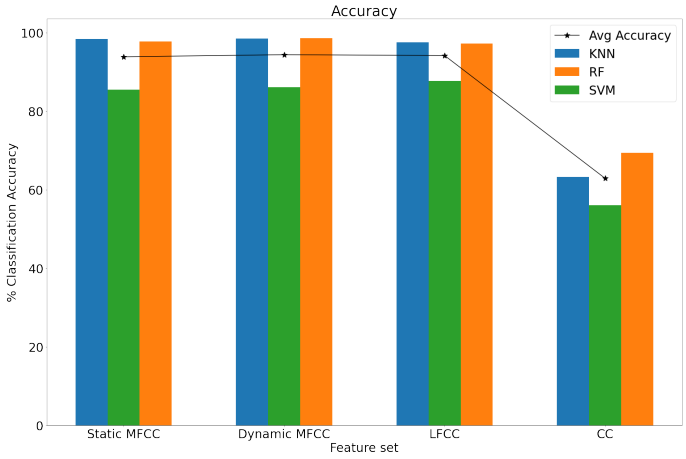

Figure 9.2: Accuracy of features with window size 55 ms. After [67].

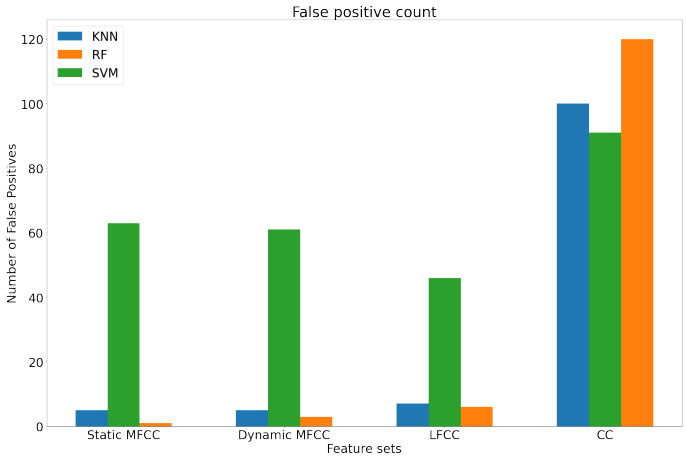

Figure 9.3: False positive (FP) of various feature vectors of window size 55 ms.

temporal axis of the matrix generated by the feature extraction technique was averaged, resulting in a conversion to a *1*-D vector. The obtained results demonstrate that there is minimal loss of information, as indicated by the Repeat-stratified 10-fold accuracy of 98.48 %. This reduction in information loss also contributes to reducing the computational complexity when using these features in classifiers or deep learning architectures. Another important objective is to minimize false positive occurrences, which is particularly crucial in realistic scenarios where misclassifying pathological cries as normal is undesirable. This objective is consistently achieved when using static Mel Frequency Cepstral Coefficients (MFCC) feature vectors across all three classifiers. The most favourable outcomes are observed when employing dynamic MFCC feature extraction in combination with K-Nearest Neighbors (KNN) and Random Forest (RF) classifiers, as illustrated in Figure 9.3.2 and Table 9.4.

The outcomes demonstrate that each machine learning classifier manages re-

Table 9.4: Confusion Matrix of the Best Feature (Dynamic MFCC). After [67].

Classifiers	Classes	Normal Class	Pathology Class
KNN	Normal	**300**	4
	Pathology	5	**370**
RF	Normal	**290**	15
	Pathology	2	**380**
SVM	Normal	**260**	43
	Pathology	61	**320**

dundant information in its own unique manner. Nevertheless, it is important to acknowledge that the gradient time-averaged feature vector inherently encompasses dynamic information. This is achieved through the concatenation of static features, delta features, and delta-delta features.

9.4 Chapter Summary

In this chapter, we observed the effect of time averaging on various feature sets. The evaluation of the features is performed using various window sizes. The time-averaged features resulted in a maximum accuracy of **98.48** % which is comparable with state-of-the-art deep learning classifiers. This work indicates that the frequency plane of an infant cry contains more discriminative information than the temporal plane Furthermore, a noteworthy observation was made regarding the relative abundance of information in lower frequencies compared to higher frequencies. In the next chapter, a novel application of music-based harmonic and pitch features extracted from CQT is applied to the infant cry classification problem while considering the cry signal to be a melodic (a prosodic characteristic) signal.

CHAPTER 10
Constant-Q Based Pitch and Harmonic Features for Infant Cry Classification

10.1 Introduction

The classification of normal *vs.* pathological infant cry sample is itself a challenging problem due to the limited amount of data and the uncontrollability of the speaker. Many state-of-the-art feature sets, such as MFCC, LFCC, and CQCC have been used for this task. The MFCC is called a state-of-the-art feature set due to its high performance for the normal *vs.* pathology classification of infant cry. However, an effective representation of the *spectral* and *pitch* components of a spectrum together is not achieved leaving scope for improvement. Also, the infant cry can be considered a melodic sound implying that the fundamental frequency and timbre-based features also carry vital information. This work proposes Constant Q Harmonic Coefficients (CQHC), and Constant Q Pitch Coefficients (CQPC) extracted by the decomposition of the Constant Q Transform (CQT) spectrum for the infant cry classification. This work uses Convolutional Neural Network (CNN) as the classifier along with traditional classifiers, namely, Gaussian Mixture Models (GMM), and Support Vector Machines (SVM). The results are compared by considering the MFCC, LFCC, and CQCC feature sets as the baseline features. Additionally, the effect of the log is observed on the proposed feature sets. The results show that the feature-level fusion of CQT-based CQHC and CQPC features outperforms the baseline features by a considerable margin. All the mentioned features are evaluated by keeping a window size of 25 ms and a hop length of 10 ms, F_{min} = 100 Hz, and octave resolution of 14. All the features were extracted using librosa toolkit [47]. The Baby Chillanto dataset is used in this work. Out of the entire dataset, 80 % is used for training and the rest of the data is used for testing the model.

10.2 Proposed Features

10.2.1 Constant-Q Harmonic Coefficients (CQHC)

The logarithmic resolution in the CQT spectrogram enables harmonic frequencies to exhibit a consistent arrangement in the frequency-domain, maintaining their relative positions with respect to the fundamental frequency (F_0) in an unchanging manner [71]. As the harmonics are the spectral coefficients carrying the spectral information of the signal, they can be used in the *timbre* characterization of the signal, where timbre can be defined as the quality of the sound produced. Given the pitch can be normalized, the locations of harmonics can be obtained and their energies be calculated leading to an efficient timbre feature set.

To achieve pitch normalization, it is assumed that the CQT spectrum can be represented as a convolution of two components: a pitch-normalized spectral component and an energy-normalized pitch component. This assumption allows for compensating pitch variations and enables more accurate comparisons and analyses of musical content, as demonstrated in Eq.10.1 [71]:

$$A = B * C, \tag{10.1}$$

where A represents the CQT spectrum, B represents the pitch-normalized spectral component, C represents the energy-normalized pitch component, and from the property that the magnitude is *shift-invariant*, the spectral component can be approximated by the magnitude of the CQT spectrum. The IFFT of the above approximation gives the estimate of the spectral component as stated in Eq 10.2 [71]:

$$B = \mathcal{F}^{-1}(|\mathcal{F}(A)|), \tag{10.2}$$

where $\mathcal{F}^{-1}$ represents the inverse Fourier transform function. Given the octave resolution considered for the calculation of CQT, we can obtain the positions of harmonics in the spectral component and then extract the harmonic coefficients. The coefficients from the spectral component are obtained by [71]:

$$i = round(O_c log_2(k)), \tag{10.3}$$

$$CQHC_k = B(i), \tag{10.4}$$

where k takes the value between 1 and N_c, O_c is the octave resolution and N_c is the

number of desired coefficients. The CQHC captures the harmonics information of the speech signal embedded in the CQT spectrum. In this work, along with CQHC, additionally, logarithmic CQHC is also considered.

Algorithm 1 Python pseudo code for CQHC and CQPC feature extraction

Input: Speech signal x(n) and sampling frequency Fs
Output: cqhc_feat, cqpc_feat

```
cqt_spec ← cqt(x(n), Fs)                                  ▷ Constant Q transform
power_cqt ← power(cqt_spec, 2)                            ▷ Power spectrum of the CQT
ft_cqt ← FT(power_cqt)                                    ▷ Fourier transform of the power spectrum
absft_cqt ← abs(ft_cqt)                                   ▷ Absolute value
spect_comp ← real(ifft((absft_cqt)))                      ▷ Pitch normalized spectral component
pitch_comp ← real(ifft(ft_cqt/absft_cqt))                 ▷ Energy normalized pitch component
indices ← round(octave_resol * log(arrange(1, numcoeff + 1)))   ▷ Indices values
cqhc_feat ← spect_comp[indices, :]
cqpc_feat ← pitch_comp[indices, :]
```

10.2.2 Constant-Q Pitch Coefficients (CQPC)

The CQT spectrum decomposition also yields an energy-normalized pitch component, which means that F0, and the first few formants are encoded within the pitch component, preserving the relevant information. The pitch component is calculated as [71] :

$$C = \mathcal{F}^{-1}(e^{jarg(\mathcal{F}(A)}). \tag{10.5}$$

where $F^{-1} represents the Inverse Fourier transform. Furthermore, the coefficients for the pitch compo$

10.3 Motivation of Harmonic and Pitch Coefficients for Infant Cry

the infant cry can be considered a melodic sound or prosodic characteristics implying that the fundamental frequency (F_0) and timbre-based features also carry vital information. Figure 10.1 represents the CQT-gram analysis of normal *vs.* asphyxia *vs.* deaf cries. From Panel III of figure 10.1, it can be observed that the pitch component of a normal cry is found to have a continuous contour plot (F_0 contour), however, it is seen to be discontinuous for the pathology cry. Furthermore, it is observed that the pitch component of pathology cries occurs at higher frequencies than the pitch component of the normal infant cry. These observations

make the pitch component a vital differentiating factor for normal *vs.* pathology infant cry. However, the CQPC component does not contain the formant information. Panel II of figure 10.1 represents the spectral component of infant cries. For pathological cries, the harmonic structures are found to be smeared, when compared with the normal infant cry. Due to the pitch normalization of the harmonics, the resolution of the harmonic component decreases. This makes the harmonic component a poor choice when considered alone.

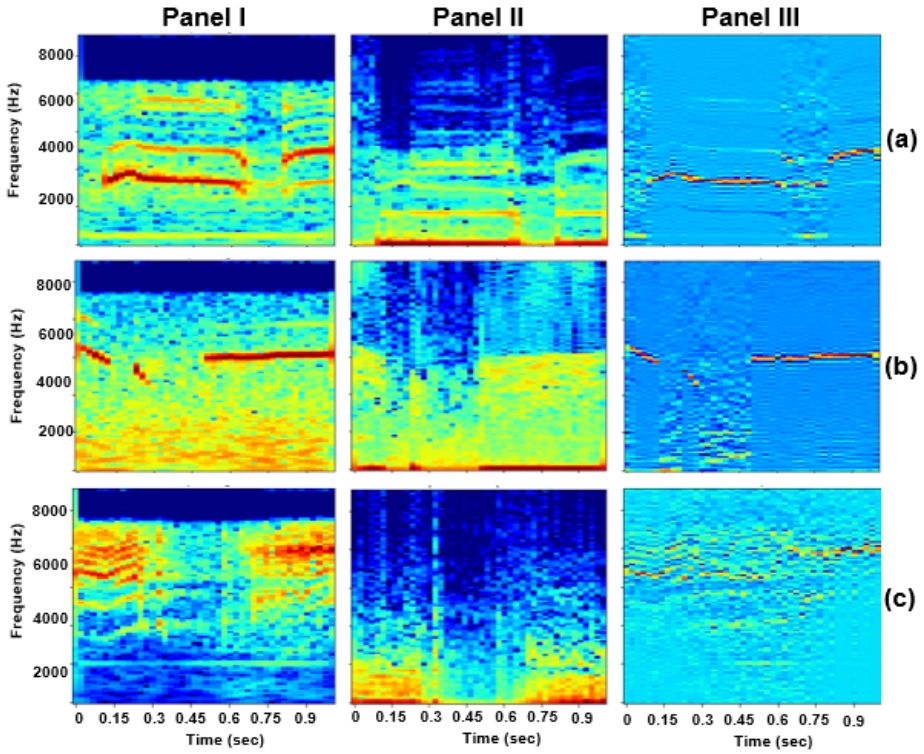

Figure 10.1: Panel I, Panel II, and Panel III Depicts CQT-gram, Spectral Component, and Pitch Component, Respectively for (a) Normal Cry, (b) Asphyxia, and (c) Deaf Cries. Best viewed in colour. After [68].

10.4 Experimental Results

10.4.1 Results for Baseline Features

The accuracy obtained using baseline features is reported in Table 10.1. It can be seen that the maximum *5*-fold accuracy of *96.95* % is achieved using the MFCC on the CNN classifier with a test accuracy of *98.24* %. Further, it can be seen that traditional classifiers such as GMM give an accuracy of 99.69 % and SVM gives an accuracy of 86.21 % for the MFCC feature set. The MFCC results in the highest

accuracy of 96.95 % as it contains generalized timbre information and pitch information [71]. It should also be noted that the introduction of the Mel scale features is primarily aimed at the musical signals [71] and since the infant cry can be considered as a melodic signal, the Mel scaled features outperforms the linear-scaled features.

Table 10.1: Results for Baseline Features for Infant Cry Classification. After [68].

Features	CNN 5-Fold Accuracy	CNN Test Accuracy	GMM	SVM
MFCC	**96.95**	**97.88**	**99.69**	**86.21**
LFCC	94.42	96.47	99.16	84.76
CQCC	93.27	93.00	95.44	83.12

10.4.2 Results for Proposed Feature Sets

The CQT feature set resulted in an accuracy of 90.32 % as shown in Table 10.2. The CQHC and CQPC feature sets, which are obtained by decomposing the CQT spectrum resulted in accuracies of 80.85 % and 83.47 %, respectively. This result indicates the importance of the pitch component for the infant cry classification, which is captured by the CQPC feature set. This might be due to the fact that the pathology cry contains irregular breathing patterns, which are caused due to affected vocal folds, and it is known that the fundamental frequency (F_0) is tied to the rate of vocal fold vibration [22]. Hence, the F_0 or the pitch component contains differential cues of the cry, which is vital for the classification task of normal vs. pathology cry. This result also indicates the fact that infant cries exhibit rich melodic features i.e., variation of fundamental frequency w.r.t time [93]. On the other hand, the CQHC feature set, which is extracted by normalizing the spectrum w.r.t F_0 fails to perform when compared with the CQPC feature set indicating the timbre information alone does not carry differentiating factors for the normal and pathological infant cry. However, neither the harmonics component nor the pitch component alone is resulting in accuracy higher than the CQT feature set. These results can be supported by the spectrographic analysis performed in the previous subsection. Furthermore, the effect of the logarithm applied to the feature sets was investigated. The application of a logarithm on any spectrum helps to increase the resolution of the spectrum. It can be observed from Table 10.2 that the effect of the log is negligible in the case of CQPC as the increase of resolution of the energy normalized pitch component doesn't add much information compared to the spectral component which contains the information of the harmonics,

that is normalized to the lowest frequency. Similar conclusions can be drawn from the results obtained using traditional classifiers. The SVM is the least-performing classifier which might be because of its inability to deal with mapping features that are not linearly separable in lower dimensional feature space into linearly separable higher dimensional feature space, where they become linearly separable.

Table 10.2: Accuracy of CQT, CQHC, and CQPC for Infant Cry Classification. After [68].

Feature Set	**CNN 5-Fold Accuracy**	**CNN Test Accuracy**	**GMM**	**SVM**
CQT	90.32	87.12	90.7	70.62
CQHC	80.85	82.00	85.77	64.49
CQPC	83.47	85.18	89.6	62.59
Log CQHC	90.70	91.12	90.22	77.27
Log CQPC	**91.24**	**92.12**	**93.61**	**80.31**

10.4.3 Results for Feature-Level Fusion of Various Feature Sets

This sub-Section discusses the results obtained from the feature-level fusion of MFCC, CQHC, CQT, and CQPC feature sets. This fusion is the concatenation of various feature sets extracted in different ways into a single feature set. The fusion of CQHC and CQPC outperforms the CQT feature set indicating that providing the pitch component separately results in a better performance. This result states that feeding the F_0 contour information separately along with harmonic information results in a better accuracy as can be seen from Figure 10.1. The addition of log to the fusion of CQHC and CQPC performs comparably with the MFCC feature and outperforms LFCC and CQCC features. MFCC manages to capture generalized timbre information in it along with the pitch information [71]. The infant cry can be considered a melodic sound due to the continuous variations in the pitch of the cry. The timbre information provides the colour for the melodic sounds. Hence, both CQHC and MFCC capture the timbre information in a unique way. The fusion of MFCC and log-CQHC features beat the baseline MFCC feature set by **1.78 %**. This indicates that the harmonic features of the CQT spectrum carry additional information when compared with the generalized harmonic features captured by MFCC. Furthermore, the feature-level fusion of MFCC and log-CQPC feature set results in an improvement in the accuracy of **2.01%**, when compared with baseline MFCC features indicating that the additional pitch information is important in the infant cry classification task. The fusion of MFCC with log CQPC

and the fusion of MFCC with log-CQPC and log-CQHC managed to beat the baseline MFCC resulting in an absolute improvement of **3 %** in accuracy. This shows that the CQHC consists of unique information obtained from the CQT spectrum, which the MFCC fails to capture. It also indicates the inability of the MFCC feature set to capture the pitch component, when compared to the CQPC feature set. Hence fusion of the MFCC feature set with the CQT decomposed features (CQHC and CQPC) resulted in a noticeable amount of increase in accuracy, which can also be observed in the traditional classifiers.

Table 10.3: Accuracy of Various Feature-Level Fusion of Features for Infant Cry Classification. After [68].

Feature Set	**CNN 5-Fold Accuracy**	**CNN Test Accurcy**	**GMM**	**SVM**
CQHC+CQPC	91.92	94.17	93.26	70.66
Log CQHC+ Log CQPC	95.35	94.70	97.53	84.32
MFCC+Log CQHC	98.73	99.29	99.34	89.91
MFCC+Log CQPC	98.96	99.47	99.52	91.63
MFCC+Log CQT	98.45	99.47	99.52	86.48
MFCC+LOG CQHC+ LOG CQPC	**99.12**	**99.47**	**98.81**	**92.47**

10.4.4 Statistical Analysis of Proposed Features

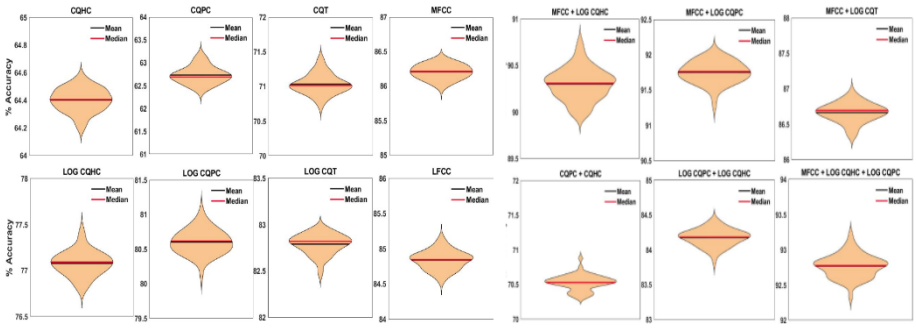

Figure 10.2: Analysis of statistical significance via violin plots for various feature sets. After [68].

The statistical significance of results is shown using stratified k-fold cross-validation to ensure similar data distribution in each fold. 5-fold CV is performed 50 times to get violin plots as shown in Fig. 10.2, which shows relatively higher mean and median than the existing features for the proposed features. It is observed that the mean and median are relatively higher for the proposed features indicating the statistical significance of the results.

10.5 Chapter Summary

This chapter discusses the importance of harmonic and pitch information in the infant cry signal. A unique approach is introduced in this study, where the application of CQHC and CQPC features is proposed for classifying infant cries. It is observed that the pitch component extracted through CQPC feature outperforms the baseline features (MFCC and LFCC). Furthermore, the effect of the logarithm function on the feature sets is observed. Finally, the statistical significance is shown through the violin plots. The next chapter concludes this thesis work and provides a brief summary of the work covered, and potential future research directions.

CHAPTER 11
Summary and Conclusion

11.1 Summary

This thesis is focused on the development of effective signal processing techniques and features for various speech-based problems, such as dysarthric severity-levels classification, infant cry classification, emotion recognition, and voice liveness detection. Furthermore, the thesis delved into the applicability of CQHC and CQPC features for infant cry classification, as well as the effect of time-averaged features for the same purpose. The utilization of phase-based group delay features proved to be effective in classifying dysarthria severity-levels, recognizing emotions, and detecting voice liveness. These features captured important temporal characteristics of speech and vocalizations, which describe the speech production system, allowing for accurate and meaningful classification across different tasks. The proposed feature techniques are evaluated using various machine learning and deep learning classifiers, such as K-Nearest Neighbor Classifier, Random Forest Classifier, Support Vector Classifier, Gaussian Mixture Models, and Convolutional Neural Network. Furthermore, the thesis evaluated the performance of the classification models using metrics, such as accuracy, precision, recall, and F1 score. The application of 5-fold cross-validation ensured a robust evaluation process, providing reliable and unbiased measures of the model's performance. These metrics served as indicators of the model's accuracy and effectiveness, aiding in the assessment of its suitability for real-world applications.

Furthermore, the thesis explored the application of CQHC and CQPC features for infant cry classification. These features offered unique insights, when the infant cry is considered a melodic signal. The importance of pitch features is observed for the infant cry problem. Additionally, the investigation of time-averaged features for infant cry classification provided valuable insights into how little amount of information is captured along the temporal-axis. It is observed that the time averaging resulted in a minimal information loss and enabled to use

of less computationally expensive machine learning classifiers while still able to achieve high accuracies.

11.2 Limitations of Current Work

Although the proposed features resulted in a remarkable accuracy results across all the problem statements considered in the thesis work, the following are some limitations of the proposed work:

- Even though the cross-database analysis provides insights into the speaker independency of the model, the Leave One Speaker Out (LOSO) technique on a single dataset provides us with a better understanding of the speaker independency of the proposed feature set.
- Since the dataset used for emotion recognition is EMODB, which is based on the German language, the results might not be generalized for other languages.
- Due to computational limitations, the experimentation for VLD is performed only on a part of the entire dataset.
- The limited amount of speech samples and the limited categories available for pathology cry is always an issue.
- Cross-database evaluation is not performed for the infant cry classification problem due to a large imbalance among the available datasets.

11.3 Future Research Directions

- To overcome the challenge of limited datasets for the large deep learning classifiers, various data augmentation techniques can be explored on the available datasets to generate realistic speech samples. Data augmentation can be explored using acoustic parameters and deep learning generative models.
- Since dysarthric speech and emotional speech contains vital temporal information, the use of sequential deep learning classifiers might result in the improvement of the performance.
- A cross-database evaluation for infant cry classification and emotion recognition among the balanced datasets.

- This work showcases the dysarthric speech analysis for the adult speaker. The same experimentation can be repeated for children speech disordered datasets.
- Dysarthric speech enhancement can be explored by reconstructing the phase as similar to that of the control speaker phase structure.
- The group delay features for the entire POCO dataset is yet to be explored. Furthermore, the phase-based features for various spoof attacks in Automatic Speaker Verification (ASV) can also be considered.
- The comparison between signal processing-based, acoustic-based features, and deep learning features learned through the transfer learning method for all the above problem statements might lead to an interesting study.

List of Publications

- **Journal Paper**

1. **Aditya Pusuluri**, Aastha Kachhi, and Hemant A Patil: "Modified Phase Based Function for Dysarthric Severity-Level Analysis and Classification" **Article under preparation** in IEEE/ACM, Audio Speech and Language Processing

- **Conference Papers**

1. **Aditya Pusuluri**, Aastha Kachhi, and Hemant A Patil: "Analysis of Time-Averaged Feature Extraction Techniques on Infant Cry Classification" **Published** in 24^{th} International Conference on Speech and Computer (SPECOM) 2022.

2. S. Uthiraa, **Aditya Pusuluri**, Hemant Patil, "Modified Group Delay Features for Emotion Recognition," **submitted** in International Conference on Pattern Recognition and Machine Intelligence (PREMI), December 12-15, Kolkata, India.

3. **Aditya Pusuluri**, Aastha Kachhi, and Hemant A Patil: "Constant-Q Based Harmonic and Pitch Features for Normal *vs* Pathological Infant Cry Classification" **Rejected** in 31^{st} European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4-8 September 2023.

4. **Aditya Pusuluri** and Hemant A Patil: "Modified Group Delay Based Cepstral Features for Dysarthria Severity-Level Classification" **Rejected** in 31^{st} European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4-8 September 2023.

5. S. Uthiraa, Akshat Vora, Prathamesh Bonde, **Aditya Pusuluri**, Hemant A. Patil, "Spectral and Pitch Components of CQT Spectrum for Emotion Recognition," **Rejected** in 31^{st} European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4-8 September 2023.

6. **Aditya Pusuluri**, Aastha Kachhi, and Hemant A Patil: "Constant-Q Based Harmonic and Pitch Features for Normal *vs* Pathological Infant Cry Classification" **Rejected** in International Conference on Acoustics, Speech and Signal Processing, Greece 4-9 June 2023.

7. **Aditya Pusuluri** and Hemant A Patil: "Noise Robustness of MGDCC for Dysarthric severity-level Classification" **Rejected** in INTERSPEECH 2023.

8. **Aditya Pusuluri**, and Hemant A Patil: "Analysis of Time-Averaged Feature Extraction Techniques on Dysarthric Severity-Level Classification" **Rejected** in ISCLP 2022

9. S. Uthiraa, Akshat Vora, Prathamesh Bonde, **Aditya Pusuluri**, Hemant A. Patil, "Combining Features from Spectral and Pitch Components of CQT Spectrum for Emotion Recognition," **rejected** in International Conference on Acoustics, Speech and Signal Processing, Greece 4-9 June 2023.

References

[1] P. Aarabi, G. Shi, M. M. Shanechi, and A. S. Rabi. *Phase-Based Speech Processing*. World Scientific Publishing Company, December, 2005.

[2] K. Akimoto, S. P. Liew, S. Mishima, R. Mizushima, and K. A. Lee. POCO: A voice spoofing and liveness detection corpus based on pop noise. In *INTERSPEECH*, pages 1081–1085, Shanghai, China, October 2020.

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

[4] D. Ap. Maximum likelihood from incomplete data via em algorithm. *J. Royal Stat. Soc. B.*, 39(1):1–38, 1977.

[5] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin. A review on emotion recognition using speech. In *2017 International conference on inventive communication and computational technologies (ICICCT)*, pages 109–114. IEEE, 2017.

[6] C. Bhat, B. Vachhani, and S. K. Kopparapu. Automatic assessment of dysarthria severity-level using audio descriptors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA*, pages 5070–5074, 2017.

[7] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine learning*, volume 4. Springer, 2006.

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

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al. A database of german emotional speech. In *INTERSPEECH, Lisbon, Portugal*, volume 5, pages 1517–1520, 2005.

[10] 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.

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

[12] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial intelligence in medicine*, 37(1):7–18, 2006.

[13] S. P. Dewi, A. L. Prasasti, and B. Irawan. Analysis of LFCC feature extraction in baby crying classification using KNN. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 86–91. IEEE, 2019.

[14] S. P. Dewi, A. L. Prasasti, and B. Irawan. The study of baby crying analysis using MFCC and LFCC in different classification methods. In *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, pages 18–23. IEEE, 2019.

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

[16] T. Drugman, T. Dubuisson, and T. Dutoit. Phase-based information for voice pathology detection. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4612–4615. IEEE, 2011.

[17] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

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

[19] P. Enderby. Disorders of communication: dysarthria. *Handbook of clinical neurology*, 110:273–281, 2013.

[20] 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.

[21] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad. Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks. *ETRI Journal*, 40(5):643–652, 2018.

[22] D. R. Feinberg, B. C. Jones, A. C. Little, D. M. Burt, and D. I. Perrett. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal behaviour*, 69(3):561–568, 2005.

[23] 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), Osijek, Croatia*, pages 181–186, 2019.

[24] D. B. Freed. *Motor Speech Disorders: Diagnosis and Treatment*. Plural Publishing, 2018.

[25] J. O. Garcia and C. R. Garcia. Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feedforward neural networks. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 4, pages 3140–3145. IEEE, 2003.

[26] J. Garland. No language but a cry, 1972.

[27] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel. Cross-database models for the classification of dysarthria presence. In *INTERSPEECH, Stockholm, Sweden*, pages 3127–31, 2017.

[28] M. A. Grudin. On internal representations in face recognition systems. *Pattern recognition*, 33(7):1161–1177, 2000.

[29] S. Gupta *et al.* Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117, 2021.

[30] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *INTERSPEECH*, pages 930–934, Lyon, France, August 2013.

[31] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Significance of the modified group delay feature in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):190–202, 2006.

[32] Hemant A. Patil. Cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry. In *A. Neustein (Ed.) Advances in Speech Recognition, Springer*, pages 323–348. 2010.

[33] M. A. Hossan, S. Memon, and M. A. Gregory. A novel approach for MFCC feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–5, 2010. Accessed: 2022-08-11.

[34] A. K. Jain, K. Nandakumar, and A. Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016.

[35] 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.

[36] A. A. Joshy and R. Rajan. Automated dysarthria severity classification using deep learning frameworks. In *2020 28th European Signal Processing Conference (EUSIPCO), Dublin, Ireland*, pages 116–120, 2021.

[37] A. Kachhi, A. Therattil, A. T. Patil, H. B. Sailor, and H. A. Patil. Teager energy cepstral coefficients for classification of dysarthric speech severity-level. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand*, pages 1462–1468. IEEE, 2022.

[38] R. D. Kent, J. F. Kent, J. R. Duffy, J. E. Thomas, G. Weismer, and S. Stuntebeck. Ataxic dysarthria. *Journal of Speech, Language, and Hearing Research*, 43(5):1275–1289, 2000.

[39] M. J. Kim, J. Yoo, and H. Kim. Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models. In *Interspeech*, pages 3622–3626, 2013.

[40] Y. Kim, R. D. Kent, and G. Weismer. An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech-Language and Hearing Research (JSLH), 54(2):417-29*, 2011.

[41] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117, 2012.

[42] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou. The importance of phase on voice quality assessment. In *INTERSPEECH, Singapore*, 2014.

[43] A. Y. Kuznetsov, R. A. Murtazin, I. M. Garipov, E. A. Fedorov, A. V. Kholodenina, and A. A. Vorobeva. Methods of countering speech synthesis attacks on voice biometric systems in banking. *Scientific and Technical Journal of Information Technologies Mechanics and Optics*, 21(1):109–117, 2021.

[44] N. Lévêque, A. Slis, L. Lancia, G. Bruneteau, and C. Fougeron. Acoustic change over time in spastic and/or flaccid dysarthria in motor neuron diseases. *Journal of Speech, Language, and Hearing Research*, 65(5):1767–1783, 2022.

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

[46] 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.

[47] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25, 2015.

[48] 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.

[49] A. Messaoud and C. Tadj. A cry-based babies identification system. In *Image and Signal Processing:* 4^{th} *International Conference, ICISP 2010, Trois-Rivières, QC, Canada, June 30-July 2, 2010. Proceedings 4*, pages 192–199,. Springer, 2010.

[50] R. P. Michelson. The results of electrical stimulation of the cochlea in human sensory deafness. *Annals of Otology, Rhinology & Laryngology*, 80(6):914–919, 1971.

[51] S. Mochizuki, S. Shiota, and H. Kiya. Voice livness detection based on pop-noise detector with phoneme information for speaker verification. *The Journal of the Acoustical Society of America (JASA)*, 140(4):3060–3060, 2016.

[52] S. Mochizuki, S. Shiota, and H. Kiya. Voice liveness detection using phoneme-based pop-noise detector for speaker verifcation. In *Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 233–239, Les Sables d'Olonne, France, 2018.

[53] H. A. Murthy and B. Yegnanarayana. Formant extraction from group delay function. *speech communication*, 10(3):209–221, 1991.

[54] H. A. Murthy and B. Yegnanarayana. Speech processing using group delay functions. *Signal Processing*, 22(3):259–267, 1991.

[55] H. A. Murthy and B. Yegnanarayana. Group delay functions and its applications in speech technology. *Sadhana*, 36:745–782, 2011.

[56] K. M. Murthy and B. Yegnanarayana. Effectiveness of representation of signals through group delay functions. *Signal Processing*, 17(2):141–150, 1989.

[57] R. Naika. An overview of automatic speaker verification system. In *Intelligent Computing and Information and Communication: Proceedings of* 2^{nd} *International Conference, (ICICC), Pune, India*, pages 603–610. Springer, 2018.

[58] R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the seventh ACM International Conference on Multimedia (Part 1)*, pages 343–351, 1999.

[59] N. Narendra and P. Alku. Dysarthric speech classification using glottal features computed from non-words, words, and sentences. In *INTERSPEECH, Hyderabad, India*, pages 3403–3407, 2018.

[60] 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.

[61] 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 {Last Accessed: 01-Feb-2022}.

[62] 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.

[63] S. H. K. Parthasarathi, P. Rajan, and H. A. Murthy. Robustness of group delay representations for noisy speech signals. Technical report, IDIAP, Switzerland, 2011.

[64] H. A. Patil. Infant identification from their cry. In *Seventh International Conference on Advances in Pattern Recognition (ICAPR), Kolkata, India*, pages 107–110. IEEE, 2009.

[65] H. A. Patil, A. T. Patil, and A. Kachhi. Constant Q Cepstral Coefficients for classification of normal vs. pathological infant cry. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore*, pages 7392–7396, 2022.

[66] A. Paul, R. K. Das, R. Sinha, and S. M. Prasanna. Countermeasure to handle replay attacks in practical speaker verification systems. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5, IISc, Bengaluru, India, June 2016.

[67] A. Pusuluri, A. Kachhi, and H. A. Patil. Analysis of time-averaged feature extraction techniques on infant cry classification. In *Speech and Computer: 24th International Conference, SPECOM 2022, Gurugram, India, November 14–16, 2022, Proceedings*, pages 590–603. Springer, 2022.

[68] A. Pusuluri, A. Kachhi, and H. A. Patil. Constant-q based harmonic and pitch features for normal vs pathological infant cry classification. In *rejected in European Signal Processing Conference (EUSIPCO), Helsinki, Finland*, 4-8 September 2023.

[69] A. Pusuluri, A. Kachhi, and H. A. Patil. Modified group delay based cepstral features for dysarthria severity-level classification. In *rejected in European Signal Processing Conference (EUSIPCO), Helsinki, Finland*, 4-8 September 2023.

[70] T. F. Quatieri. *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2002.

[71] Z. Rafii. The Constant-Q Harmonic Coefficients: A timbre feature designed for music signals [Lecture Notes]. *IEEE Signal Processing Magazine*, 39(3):90–96, 2022.

[72] P. Rajan, T. H. Kinnunen, J. Pohjalainen, P. Alku, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, et al. Using group delay functions from all-pole models for speaker recognition. INTERSPEECH, Lyon, France, 2013.

[73] R. A. Rashid, N. H. Mahalin, M. A. Sarijari, and A. A. A. Aziz. Security system using biometric technology: Design and implementation of voice recognition system (vrs). In *2008 International Conference on Computer and Communication Engineering*, pages 898–902, 2008.

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

[75] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, and E. Arch-Tirado. Infant cry classification using genetic selection of a fuzzy model. In *Iberoamerican Congress on Pattern Recognition*, pages 212–219. Springer, 2012.

[76] 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.

[77] A. E. Rosenberg. Automatic speaker verification: A review. *Proceedings of the IEEE*, 64(4):475–487, 1976.

[78] F. Rudzicz. Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960, 2010.

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

[80] C. P. Sellors. Carl plantinga and greg m. smith (eds.), Passionate Views: Film, Cognition, and Emotion, 2000.

[81] N. Shah and P. Shrinath. Iris recognition system–a review. *International Journal of Computer and Information Technology*, 3(02):321–327, 2014.

[82] 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), Kanpur, India*, pages 1–6, 2019.

[83] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui. Voice liveness detection algorithms based on pop noise

caused by human breath for automatic speaker verification. In *INTERSPEECH,Dresden,Germany*, pages 239–243, 2015.

[84] Shiota, Sayaka and Villavicencio, Fernando and Yamagishi, Junichi and Ono, Nobutaka and Echizen, Isao and Matsui, Tomoko. Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector. In *Speaker Odyssey, Bilbao, Spain*, volume 2016, pages 259–263, 2016.

[85] M. Swain, A. Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[86] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan. Emotion in a century: A review of emotion recognition. In *proceedings of the 10th international conference on advances in information technology*, pages 1–8, 2018.

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

[88] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth. A multitask learning approach to assess the dysarthria severity in patients with parkinson's disease. In *INTERSPEECH, Hyderabad, India*, pages 456–460, 2018.

[89] K. Vijayan, P. R. Reddy, and K. S. R. Murty. Significance of analytic phase of speech signals in speaker verification. *Speech Communication*, 81:54–71, 2016.

[90] E. Vyzas. *Recognition of emotional and cognitive states using physiological data*. PhD thesis, Massachusetts Institute of Technology, USA, 1999.

[91] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2062–2070, 2019.

[92] 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.

[93] K. Wermke and W. Mende. Musical elements in human infants' cries: in the beginning is the melody. *Musicae Scientiae*, 13(2_suppl):151–175, 2009.

[94] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66:130–153, 2015.

[95] 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.

[96] B. Yegnanarayana, D. Saikia, and T. Krishnan. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(3):610–623, 1984.

[97] K. M. Yorkston, D. R. Beukelman, and C. Traynor. *Assessment of Intelligibility of Dysarthric Speech*. Pro-ed Austin, TX, 1984.

[98] 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.

[99] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda. Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*, 2020, Last Accessed Date=23rd May 2023.

[100] W. Ziegler and D. von Cramon. Spastic dysarthria after acquired brain injury: An acoustic study. *International Journal of Language & Communication Disorders*, 21(2):173–187, 1986.