Significance of Teager Energy Operator for Speech Applications

by

ANAND SAJU THERATTIL
202015005

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

MASTER OF TECHNOLOGY
in
ELECTRONICS AND COMMUNICATION

with specialization in
Wireless Communication and Embedded Systems
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

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

May 2022

Declaration

I hereby declare that

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

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

Anand Saju Therattil

Certificate

This is to certify that the thesis work entitled, “Significance of Teager Energy Operator for Speech Applications” has been carried out by Mr. Anand Saju Therattil for the degree of Master of Technology in Electronics and Communications at *Dhirubhai Ambani Institute of Information and Communication Technology & C.R.Rao Advanced Institute of Applied Mathematics, Statistics and Computer Science* under our supervision.

Prof. (Dr.) Hemant A. Patil
Thesis Supervisor

Acknowledgments

“The weight I must carry has become insupportable, the labour before me is now insufferable, so I carried it along until I became indestructible”. First and foremost, I am greatly thankful to my supervisor, Prof. Hemant A. Patil. He has constantly motivated and encouraged me in doing quality research through helpful discussions, manuscript corrections, and appreciating the efforts by us even when we fail several times. I would like to appreciate the enormous patience shown by sir, which was tested more than once. I would also like to thank Speech Research Lab, DA-IICT, for nurturing and strengthening emerging young minds like me in becoming future researcher of the world. I would also like to thank DA-IICT Gandhinagar, for making my stay as comfortable as possible from the 2^{nd} wave of COVID-19 pandemic and providing me with all the necessary resources, such as books, sports, green, and clam environment for research work.

I would like to thank PhD scholars Mr. Ankur T. Patil and Ms. Priyanka Gupta for their constant efforts, patience, and suggestions they have shown towards guiding me in my thesis. I would sincerely thank the Speech Research Lab alumni Dr. Hardik B. Sailor, and Dr. Madhu R. Kamble for their timely suggestion and support in my research work. I would like to thank my friends Ms. Aastha Kachhi, Ms. Shreya Chaturvedi, and Mr. Piyush Chodingala, who kept me motivated and encouraged during my thesis journey. I would also thank Mr. Lalit Kumar, Mr. Teja Vardan Reddy, Mr. Durgesh Mishra, Ms. Ami Pandat, Mr. Rishab Pandit, Mr. Sachin Parekh, Mr. Aditya Zala, Mr. Naiteek Jasani, Ms. Megha Gupta, Mr. Vishvesh Panchal, and Prof. Pritam Anand for supporting me mentally and physically during the Covid-19 pandemic. Finally, I would wholeheartedly express my gratitude to my family members for providing me with all the support and encouragement throughout my study and research and this thesis writing. Lastly, I would like to thank all those whose name have not been mentioned, however, were present during my studies.

Contents

Abstract

Speech is used in various applications apart from voice communications, such as pathology detection, severity-level classification of dysarthria, and replay spoof speech detection for voice biometric and voice assistants. The first part of this thesis work deals with the development of the countermeasure (CM) system for replay Spoof Speech Detection (SSD). Replay attack on voice biometric, refers to the fraudulent attempt made by an imposter to spoof another person's identity by replaying the pre-recorded voice samples in front of an Automatic Speaker Verification (ASV) system or Voice Assistants (VAs). Lastly, the dysarthria, which is neuromotor speech disorder is studied and analysed using various speech processing and deep learning approaches.

Dysarthria, Parkinson's disease, Cerebral Palsy, etc. are types of atypical speech, which impairs neuromotor functions of the human body. Among these, dysarthria is one of the most common atypical speech. To analyse the dysarthic condition of the patient depends on the severity-level, which is generally provided by Speech Language Pathologist (SLPs). However, to make the assessment immune to human biases and errors, this thesis is oriented towards developing the severity-level classification system using signal processing and deep learning approaches for dysarthric speech. This presents analysis of dysarthic *vs.* normal speech using the Teager Energy Operator (TEO)-based Teager Energy Cepstral Coefficients (TECC), and Squared Energy Operator (SEO)-based Squared Energy Cepstral Coefficients (SECC) as the frontend features. These features provided as input for deep learning and pattern recognition model predicts the severity-level class for dysarthria.

Lastly, the generalization of the countermeasure system for the replay attacks on the ASV systems and VAs is analysed using the TEO-based TECC feature set. The generalization of the CM system is presented through the cross-database evaluation between the Voice Spoofing Detection Corpus (VSDC), ASVspoof 2017 version 2.0 and ASVspoof 2019 PA datasets. Further, the analysis of One-point Replay (1PR) and Two-Point Replay (2PR) are presented in this thesis.

Keywords: *Replay Spoof Speech, Teager Energy Cepstral Coefficients, Cross-Teager Energy Cepstral Coefficients, Squared Energy Cepstral Coefficients, Dysarthria*

List of Principal Symbols and Acronyms

1PR	One-Point Replay
2PR	Two-Point Replay
ASV	Automatic Speaker Verification
CM	Countermeasure
CNN	Convolutional Neural Network
CQCC	Constant-Q Cepstral Coefficients
EER	Equal Error Rate
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
LCNN	Light Convolutional Neural Network
LDA	Linear Discriminant Analysis
PSD	Power Spectral Density
ReMASC	Replay Attack Microphone Array Speech Corpus
ResNet	Residual Neural Network
SECC	Squared Energy Cepstral Coefficients
SEO	Squared Energy Operator
SS	Synthtic Speech
STFT	Short-Time Fourier Transform
TECC	Teager Energy Cepstral Coefficients
TEO	Teager Energy Operator

UA-Speech Corpus	Universal Access Speech Corpus
VA	Voice Assistant
VC	Voice Conversion
VSDC	Voice Spoofing Detection Corpus

List of Tables

List of Figures

CHAPTER 1
Introduction

Communication between humans is primarily and widely performed through speech. It is a type of quasi-periodic signal with the primary goal of conveying information from one person to another, or machine to person, and vice-versa [74]. However, speech is of dynamic and complex nature as speech production mechanism involves various sophisticated biological systems, such as lungs, vocal tract, larynx, mouth, tongue, and lips [74]. The complexity of the speech signal can be further observed through the changes in the duration of pauses in the speech wave. There are variety of information, which can be captured through a normal speech such as linguistic information, gender, attribute, emotion, acoustical environment information, and health [75]. The advancement of the severity-level of patient's ailment, such as dysarthria, Parkinson's disease, cerebral palsy, apraxia, etc. can be investigated through atypical speech [59]. Furthermore, the speech processing is applied in the detection of the pathology in infants through their cry modes [26, 82]. Apart from the speech pathology, speech processing is utilized in automatic speaker recognition system using the speaker-specific information for the Speaker Identification (SID) and Automatic Speaker Verification (ASV). A SID system recognize an individual speaker from a group of speakers [66]. An ASV system, on the other hand, authenticates or verifies the claimed speaker identity from the speech utterance of the speaker [78].

The ASV system uses various biometric attributes, such as voice, fingerprint, palm print, veins, iris, face, etc. Amongst these biometrics, face, voice, and iris are the most commonly used biometrics [39]. Naturally, humans have an ability to identify a person jointly using face and voice biometrics. However, due to the recent advancement in speech technologies, voice biometrics are widely used in ASV systems. Additionally, the growth of speech technologies has made advancement in the medical-domain, where detection of the patients' ailment and severity-level can be detected through the patient's speech. Further, the advantage of use of speech technologies in voice biometric and pathology detection is due to its simplicity and touch-free interface provided by the systems, such as ASV systems, SID systems, Voice Assistants (VAs) and pathology detection. Hence, the speech technologies provide convenience and efficiency in various applications.

1.1 Motivation

The speaker verification technologies use speaker-specific attributes known as *features,* which are generated using signal processing and deep learning approaches for the ASV systems. The modern ASV system uses the advanced machine learning and deep learning approaches for the robust and efficient performance. However, the advancement in computational capabilities has increased the vulnerability of the spoofing attack through the machine learning and deep leaning algorithms [39]. Spoofing attacks indicates occurs when an imposters masquerade as a genuine speaker using the various spoofing techniques. Among the various spoofing attacks, Voice Conversion (VC), Synthetic Speech (SS), impersonation, twins, and replay attacks are the known spoofing attacks [18, 24, 63, 96]. Among these attacks, the replay attack is the easiest to implement but difficult to detect as it uses ready-to-use recording devices, such as simple tape recorder and smartphone, where requirement of technical expertise is minimum. In replay attack, an unauthorized attacker try to gain the access to a secure content through the playback of genuine speech. Hence, to protect the speech operated devices, such as ASV system and Voice Assistants (VAs), the development of Countermeasure (CM) system is essential.

To that effect, various standard datasets, such as ASVSpoof 2015, ASVSpoof 2017, ASVSpoof 2019, and ASVSpoof 2021 are released for designing CM system for ASV systems and subsequently, Replay Attack Microphone Array Speech Corpus (ReMASC) and Voice Spoofing Detection Corpus (VSDC) are released for designing CM for VAs. The differences between the CM system of ASV system and VAs exists [4]. Given the significant difference between CM systems of ASV system, the generalization of CM system is necessary.

Secondly, dysarthria is one of the most common speech disorder, occurring due to neurological diseases and neuro-degenerative disorder reported by the America Speech-Language-Hearing Association (ASHA) [5]. Dysarthria prominently occurs in the patients of Parkinson's Disease (PD) [79]. Apart from PD, dysarthria also occurs in neurological diseases and injuries, such as Cerebral Palsy, muscular dystrophy, Amyotrophic Lateral Sclerosis (ALS), brain stroke, etc. Hence, due to the growth of speech technologies has made it feasible for the detection of pathology, in underdeveloped countries, where the numbers of medical practitioners are less. Moreover, people suffering from dysarthria have significant difference characteristics than normal speech, which makes it difficult to use the conventional assistive technologies due to their impaired neuro-motor disorder.

Given these vulnerabilities of CM systems, the severity-level classification for dysarthria and the detection of pathology in infant cry, there exists a need of further exploration in these research areas.

1.2 Research Problems

1.2.1 Generalization of CM System

Self-classification is a widely used evaluation procedure, where the training and evaluation of the model is performed on the same database, which is generated through the same distribution, i.e., the data collection is limited and under similar scenarios. However, the generalization capabilities of features and classifiers cannot be estimated through self-classification. Hence, for the generalization of CM system, *cross-database* evaluation is necessary. In cross-database evaluation, the training and evaluation is performed on different databases. Further, cross-database evaluation presents overfitting of the classifiers and depicts practical potential of features in the real-world application.

1.2.2 Dysarthria

Dysarthria is a neuromotor speech disorder caused due to neurological damage, which hinders the speech production and perception depending on the severity-level. The patient suffering from dysarthria finds difficulty in communicating and expressing vocal emotions as it affects the dynamic motion of articulators, such as lip, teeth, tongue, throat, lips, and upper respiratory tract system. Further, as the dysarthria severity-level increases, the patient's condition deteriorates. Hence, analytical study of severity-level classification plays an important role for diagnosis and treatment of patient. However, symptoms of dysarthria depends on impact and area of neurological impact, and it varies from patient to patient, common symptoms of dysarthria are as follows [58]:

1. **Hoarse Quality of Voice :** Speech will sound breathy, raspy, strained, or will be softer in volume or lower in pitch.

2. **Articulation Problem:** Difficulty to physically produce a sound or sounds due to improper coordination between jaws, lips, tongue, etc.

3. **Less speech loudness:** Intensity or loudness of the dysarthric speech is less than natural speech.

4. **Slow and slurry speech:** Words spoken are slow or garbled.

1.3 Application Scope of this Thesis Work

1.3.1 Designing CM System

The requirement of genuine speech is of significant importance in various applications, such as Internet/telephonic banking, digital voicemails, smarthome devices, telephone industry, etc. The authenticity of the speaker is provided through the CM system, which uses various speaker-specific properties for the detection between genuine *vs.* spoof speech. Hence, to secure the user from these replay spoofing attacks, in this thesis design frontend CM system is implemented.

1.3.2 Severity-Level-based Classification

Recent study in dysarthic speech has primarily focused on classifying between normal and dysarthic speech. The classification of the dysarthic severity-level has been less explored. The severity-level classification aids in tracking the patient's advancement in the medication period. Further, for designing better dysarthric speech enhancement and ASR, the severity-level classification of dysarthria can be implemented. Hence, this thesis also focuses on the dysarthric classification based on severity-level into four severity-levels.

1.4 Contributions of the Thesis

1.4.1 Teager Energy Operator (TEO)

In this study, we have employed the TEO proposed in 1990 for capturing the nonlinear energy generated in the speech production mechanism. Hence, TEO-based Teager Energy Cepstral Coefficients (TECC) were used for the severity-level classification of dysarthria. Further, TECC is known to capture the reverberation effects of the replay, which results in the better performance of the CM system for ASV system and VAs.

1.4.2 Squared Energy Cepstral Coefficients (SECC)

Here, the feature representation of a speech signal is obtained using squared energy-based on linear filtering theory (specifically, on Parseval's energy equivalence). The performance of SECC indicates the relation of the linearity in the speech generation signal and the severity-level of dysarthria.

1.5 Organization of the Thesis

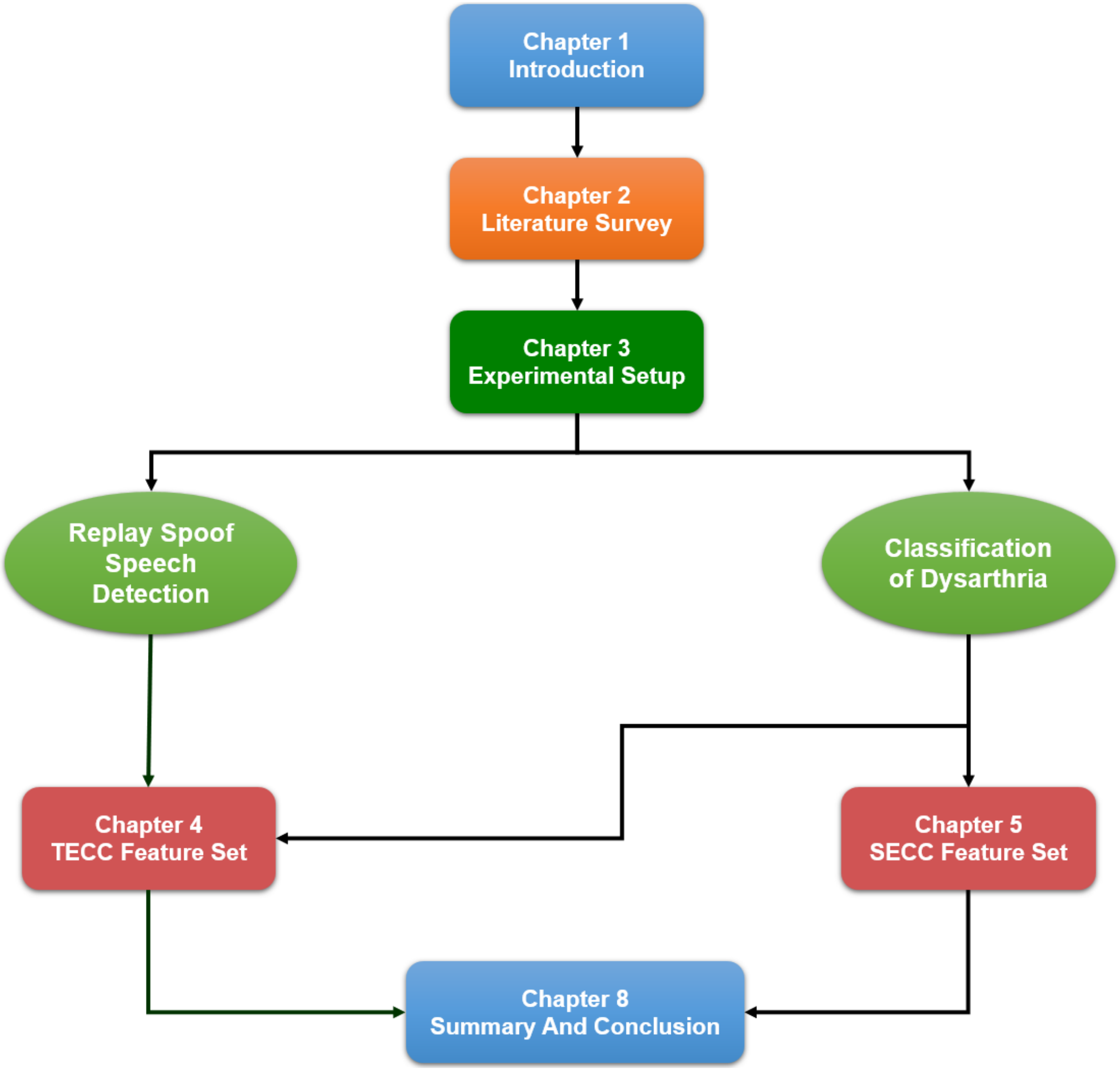

Figure 1.1: Organization of the Thesis.

- **Chapter 2** presents the detailed study on the previous investigations on replay Spoof Speech Detection (SSD), severity-level classification of dysarthria, and infant cry classification. Various methods based on signal processing and deep neural network aspects on various datasets are also discussed.

- **Chapter 3** illustrates the replay SSD, severity-level classification, and infant-cry classification methodologies and various steps required for the implementation. Further, this chapter presents the features, classifiers, and the performance measures used for the evaluation of the systems.

- **Chapter 4** uses the Teager Energy Cepstral Coefficients (TECC), which is an energy-based feature, for the designing of the CM system and severity-level classification of dysarthria. Further, TECC captures the energy generated during the speech production mechanism and hence, it is implemented for the replay SSD and dysarthric severity-level classification. Further, TECC captures the reverberation of environment, which assists in the detection of the replay SSD.

- **Chapter 5** is an extension of work on dysarthric severity-level analysis and classification based on the conclusions made in chapter 4. In this study, we proposed the Squared Energy Cepstral Coefficients (SECC) depicts the relation between the severity-level of dysarthria and presence of linearities in the dysarthic speech.

- **Chapter 6** concludes our research with a summary of the completed work within the scope of the thesis. Further, we also discuss some limitations of our work and propose some future research directions for practical potential of our thesis work.

1.6 Chapter Summary

In this chapter, we briefly looked over introduction of the CM system for replay SSD, dysarthric severity-level classification and infant cry classification as three key problem statements. Further, Section 1.1 gave an insight into the motivation that transpired into the formation of this thesis work, while Section 1.2 throws some light on the current major research issues in the scope of this thesis. Section 1.3 discusses the possible practical application of this thesis work. In Section 1.4, provides, the original contribution towards the thesis work has been described. Lastly, Section 1.5 discusses the structure and organization of the thesis. In the next chapter, we discuss the previous works and their limitations in form of literature survey in order to position this thesis work in the history of previous work.

CHAPTER 2

Literature Survey

2.1 Introduction

This chapter presents a brief literature review of a few studies that have been made in the past for replay SSD, and severity-level classification of dysarthria This chapter starts with the mathematical modelling of a speech signal through the convolution (i.e., LTI system) operation between various impulse response of the biological systems. The speech production mechanism is extended towards the mathematical modelling of the replay speech, in order to be able to design CM system effectively. In addition to replay SSD system, this chapter focuses on the dysarthric speech, the techniques used by the speech pathologist for detection and classification of dysarthria. Lastly, the modern approaches for the severity-level classification of dysarthria is discussed.

2.2 Replay Spoof Speech Detection (SSD)

2.2.1 Speech Production Mechanism

The organs involved in the speech production mechanism can be divided into three major groups: the lungs, larynx, and vocal tract system, as shown in Fig. 2.1. The lungs serve as a source of energy and provides airflow to the larynx stage of the speech production process. The larynx regulates the flow of air and generates either noisy airflow or puff-like source for the third organ group, i.e., the vocal tract system. The vocal tract system comprises oral, nasal, and pharynx cavities, these organs spectrally shape the regulated airflow provided by the larynx. Hence, the speech production mechanism can be modelled as:

$$x(t) = p(t) * l(t) * v(t), \tag{2.1}$$

where $*$ represents the convolutional operator, $p(t)$, $l(t)$, and $v(t)$ represents the source energy by lungs, airflow modulation by larynx, and transfer function of vocal tract system, respectively.

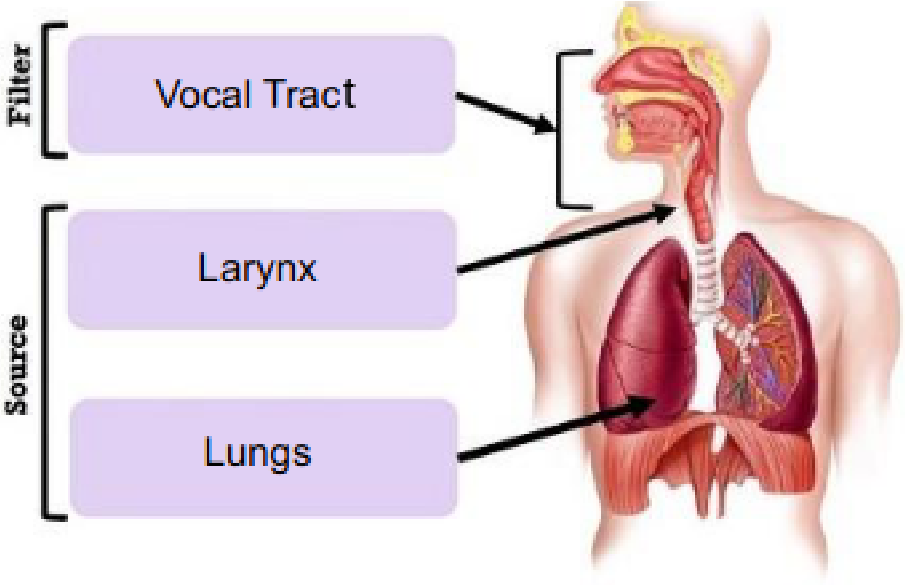

Figure 2.1: Simplified View of Speech Production Mechanism. After [50].

2.2.2 Mathematical Modelling of Replay Speech Signal

A typical replay signal may be recorded and replayed with different devices and under different environmental conditions. Hence, the mathematical model of a replay speech signal $(y(t))$ can be represented as the convolution of genuine speech $(x(t))$ with the impulse responses of recording device $(h_{rd}(t))$, recording environment $(hre(t))$, playback device $(h_{pd}(t))$, and playback environment $(h_{pe}(t))$. All of these factors lead to distortions that are introduced in the original speech signal.

$$y(t) = x(t) * h_{rd}(t) * h_{re}(t) * h_{pd}(t) * h_{pe}(t), \tag{2.2}$$

where $*$ denotes the convolution operation. This equation can also be written as:

$$y(t) = x(t) * h(t), \tag{2.3}$$

where $h(t)$ considers the cascaded effect of all the environments and intermediate replay devices.

2.2.3 Literature Review

An anti-spoofing system finds its application in safeguarding ASV system and Voice Assistants (VAs) against possible threats. Hence, to avoid any unwanted hacks, countermeasures needs to be developed. The present voice-based ASV systems are prone to spoofing attacks, namely, speech synthesis (SS) [97], voice conversion (VC) [98], replay [73], and twins [83]. Apart from the above-mentioned attacks, VAs are also vulnerable to additional attacks such as, hidden voice commands [16], self-triggered attacks [30], and audio adversarial examples [17]. In practice, we would expect our ASV system and VAs to be robust against any or all of the possible spoofing attacks.

Replay unlike any other spoofing attack is the most accessible kind of spoofing attack, wherein the attacker tries to masquerade the target simply by replaying the prerecorded voice samples [73]. The replay speech recorded with a high quality recorder and playback device in a noiseless recording environment is very hard to detect easily, as it is very equivalent to the genuine speech [73]. Hence, the present VAs and ASV systems find it very challenging to detect specially constructed signals. A replayed speech can mathematically be modelled as convolution of genuine speech signal and the impulse response of the recording device, recording environment, playback device, and the playback environment [55].

Since 2015, bi-annual ASV spoof campaigns were initiated to develop countermeasures against various spoofing attacks. Such initiatives provided standard corpora, metrics, and baseline SSD system proposal to support common evaluation. Since then, four international challenges have been organized so far to promote research in this direction, namely, ASVspoof 2015, ASVspoof 2017, ASVspoof 2019 and ASVspoof 2021. The ASVspoof 2015 challenge considered SS and VC attacks and countermeasures against these were to be developed [103]. Feature extraction methodologies for speech applications are either based on the speech production models or auditory models. The features which are derived from linear prediction (LP) analysis comes under the former class. The various LP-based features used in SSD task are LPCC, LFRCC [34, 84]. In replay SSD task, LFCC is found to be performing better than MFCC [45]. Furthermore, LFCC is chosen as feature set to develop the baselines in ASVspoof-2017 and ASVspoof-2019 challenge campaign [94, 99]. Moreover, LFCC and CQCC feature set is selected as baseline for recently released ASVspoof 2021 challenge campaign [67].

Feature set, such as Constant-Q Cepstral Coefficients (CQCC) [90], Linear Frequency Cepstral Coefficients (LFCC), and Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) [72] were a few notable contributions of this anti-spoofing challenge. The ASVspoof 2017 challenge focused on developing countermeasures against replay spoofing attacks. A comparative study of ASVspoof 2015 and ASVspoof 2017 detection challenges suggests that the detection of replay attack is more difficult as compared to the SS and VC spoofing attacks. Further, ASVspoof 2019 challenge focuses on developing countermeasures against all the three, i.e., SS, VC, and replay attacks. In addition to SS, VC, and replay attacks, DeepFake spoofing attack was introduced in recent ASVspoof 2021 challenge, generated through the deep learning framework. In addition to the ASV challenges, the Realistic Replay Attack Microphone Array Speech Corpus (ReMASC), which comprises genuine and realistic replay speech, was developed in 2019 to protect VAs from spoofing attacks. Furthermore, the Voice Spoofing Detection Corpus (VSDC) was released in 2021 for SSD development for One-Point Replay (1PR) and Two-Point Replay (2PR) threats. Playback of pre-recorded genuine utterances (0PR) to the VAs is known as 1PR attacks. Furthermore, 2PR attacks are generated through the playback of the 1PR utterance using *Drop-in* feature of the VAs.

In recent years, the CM system of ASV system and VAs were assumed to be identical, however, there are significant differences between them. For example, in ASV system, single channel microphones are present, and the user speaks generally in close proximity of the system in a close controlled acoustic environment whereas, in VAs, voice commands are spoken in diverse acoustic environment and from an unknown distance. Further, prior works on detection of replay attacks on ASV systems were based on the single channel inputs, which provided only the *spectral* and *temporal* features. However, in the VAs, microphone arrays are present, which provides spatial diversity, in addition to the spectral and temporal features which may be useful to distinguish between genuine and replay utterances. Additionally, the VAs are mounted with speech enhancement technologies, due to which the acoustic environmental noises are suppressed. Hence, due to these significant differences of ASV system *vs.* VAs, notable contributions to ASVspoof 2017, ASVspoof 2019, ReMASC, and VSDC databases were studied for the development of replay SSD shown in the Table 2.1.

Table 2.1: Selected Literature of Anit-spoofing for ASV and VAs

Database	**Author**	**Feature Sets**	**Classifier**	**% EER**	
				Dev	**Eval**
ASVspoof 2017	*Weicheng Cai et al. [14]*	CQCC (Baseline)	GMM	10.25	22.39
	Patil et al. [73]	MFCC	GMM	26.78	26.31
		CFCCIF	GMM	12.98	14.77
		LFCC	GMM	16.76	13.9
	Kamble et al. [48]	TECC	GMM	9.55	11.73
	Lian Huang et al. [104]	Hybrid Feature	GMM	8.67	25.63
		Hybrid Feature	DesnseNet	5.62	12.39
		Hybrid Feature	LSTM	9.45	15.64
		CQCC	DenseNet	7.65	17.73
		MFCC	DenseNet	6.77	15.86
		CQCC	DenseNet-LSTM	3.87	12.64
ASVspoof 2019	*Massimiliano Todisco el at. [94]*	CQCC	GMM	9.87	11.198
		LFCC	GMM	11.797	13.012
ReMASC	*Rajul Acharya el at. [71]*	CQCC	GMM	—	10.84
		Hybrid Feature	DenseNet	—	10.93
VSDC, ASVspoof 2019, ASVspoof 2017	*Anand Therattil el at. [87]*	TECC	GMM	20.01	24.07
			CNN	27.63	26.19
VSDC	Roland Baumann et al. [6]	CQCC	GMM	**0PR-1PR Test set**	**0PR-2PR Test set**
				17.29	8.78

2.3 Severity-Level Classification of Dysarthria

2.3.1 Brief History of Dysarthria Severity Classification

Dysarthric speech has complex characteristics and often requires the expertise of Speech Language Pathologists (SLPs) for diagnosis and severity-level assessment. SLPs provides severity-rating through the clear procedures of diagnosis through patient's acoustics and articulation. There are five assessment methods which are widely used by SLPs for the diagnosis of dysarthria. A brief description of these methods are discussed below:

- **Assessment of Intelligibility of Dysarthic Speech (AIDS)**

  AIDS take into consideration the intelligibility and speaking rate of a dysarthic speaker, who needs to be at least 12 years of age [100]. The voice utterance of speaker is recorded and played back to one or more examiner, who are not present at recording. The examiner then scores the speaker based on percentage intelligibility at word and sentence-level. Average scores are considered to rate the intelligibility of the speaker.

- **Speech Intelligibility Test (SIT)**

  SIT was introduced in 1996, which is an electronic form of AIDS [25]. Computer software provides the speech stimuli to the patient and scores the patient similar to AIDS.

- **Frenchay Dysarthria Assessment (FDA)**

FDA determines the type of dysarthria the patient is suffering from [15]. FDA creates a patient's profile by considering different behaviour that corresponds to speech functions, such as respiration, reflexes, intelligibility, movement of jaw, tongue, lips etc. Intelligibility is considered at word, sentence, and conversation-levels.

- **Dysarthria Examination Battery (DEB)**

  DEB evaluates the dysarthic severity-level by focusing on the prosody, articulation, phonation, resonation, and respiration [35]. Articulation is evaluated at both word and sentence-level, and the performance is evaluated by an expert diagnostician on a 5-point scale.

- **Dysarthria Profile**

  Dysarthria Profile is a 5-point rating scale and provides a comprehensive assessment of dysarthria speech [76]. The patient is rated on phonation, articulation, intelligibility, respiration etc. The ratings are provided based on one expert diagnostician and one unfamiliar listener based on reading aloud tasks.

2.3.2 Literature Review on Dysarthric Speech Classification

Although, the above methods were developed decades ago and are still widely used, many SLPs rely on informal methods for the assessment of dysarthria. A study in [35] reported that 35% SLPs use formal methods for dysarthric speech assessment. Assessment using subjective methods are also subject to familiarity bias, that is the assessment of an individual gets effected, depending on the relationship of the examiner with the patient. In addition, a study in [35] reports that there is a higher variability among the performance of the listener who transcribe atypical speech. Therefore, there is a need for better objective analysis and classification methods for dysarthic speech. Further, in recent years, through the increase of high computational systems, the development and implementation of complex deep learning architectures has been possible. Furthermore, deep learning methods provide as solution towards this problem as they are immune to human biases and errors. Therefore, recently, implementation of various signal processing combined with deep learning models has been implemented for the severity-level classification of dysarthria shown in Table 2.2.

Table 2.2: Summary of Work Done in Dysarthric Speech Classification.

Author	**Feature Sets**	**Classifier**	**Classification Type**	**% age Accuracy**
Paja et al. [70]	Multiple acoustic feature	Mahalnobis Distance	2-level severity	95
KL Kadi et al. [44]	Multiple acoustic feature	LDA+ GMM/SVM	4-level severity	93
Haewon Byeon et al. [13]	Cepstral peak prominance, jitter, shimmer etc.	Random Forest	Dysarthria *vs.* presbyphonia	83
C.Bhat et al. [8]	Audio descriptor feature	ANN	4-level severity	98
Chandrashekar et al. [20]	Mel spectrogram	CNN	3-level severity	66
M. Fernandez et al. [28]	Log-mel spectrogram	LSTM with attention	3-level severity	77
Joshy et al. [41]	MFCC	CNN	4-level severity	96
J.C. Vasquez-Correa et al. [92]	STFT spectrograms	CNN	Normal *vs.* Dysarthria	86

2.4 Chapter Summary

This chapter began with the convolutional (LTI) model of a replayed speech signal. This expression suggested that replay detection is a problem of deconvolution and hence, it is hard to detect replay. Furthermore, the history of dysarthria and the limitations of the conventional approaches is discussed. The next chapter discusses the experimental setup that is used for the replay SSD, severity-level classification and infant cry classification.

CHAPTER 3
Experimental Setup

3.1 Introduction

This chapter describes the experimental setup for the replay SSD, and severity-level classification of dysarthria. This involves description of the basic speech processing methodology, which includes pre-processing, feature extraction, and post-processing. Further, the brief description of the dataset utilized is discussed in this thesis work. In addition to the datasets and speech processing methods, various statistical and deep learning-based classifiers utilized to evaluate are discussed in the thesis work. Lastly, the performance measures are discussed for the analysis and the potential practical applications of features and classifiers.

3.2 Basic Speech Processing Methodologies

Classification system requires discriminative acoustic cues as features generated through signal processing or deep learning or combination of both the approaches for the replay SSD, severity-level classification of dysarthria, and infant cry classification. Hence, input features to classifier are extracted through the following three process: (1) pre-processing, (2) feature extraction, and (3) post-processing explained in next the sub-Section.

3.2.1 Pre-Processing

In order to extract the features effectively from the input data, it is important to convert the input data into a form which is more useful for further processing. This is done by pre-processing on the raw speech data to allow only the useful information to be used for the feature extraction process. There are several ways of pre-processing, which include pre-emphasis and silence removal schemes. However, in this work, only pre-emphasis has been used for replay SSD.

Pre-emphasis

Studies suggest that high frequency regions are more important for replay SSD as compared to the low frequency region. Hence, pre-emphasis is essentially a high pass filter that allows only the high frequencies content of the signal to pass while stopping the low-frequency content. This ensures only the part which is essential for replay SSD to be used for feature extraction. The transfer function of pre-emphasis filter is defined as [93]:

$$H(z) = 1 - \alpha z^{-1}, \tag{3.1}$$

where the value of α is 0.97 for most of the speech applications.

3.2.2 Feature Extraction

The natural speech production mechanism is a random process, and hence, speech as ensemble could not be used directly be used for classification. So the input speech signal is mapped to a vector space, which gives more effective representation of the speech signal. In this study, we have implemented the energy-based feature Teager Energy Cepstral Coefficient and Square Energy Cepstral Coefficient. Hence, these features extracted through the signal processing approach highlights the discriminative cues such as, linguistic information, acoustical environmental condition, and noise present during speech production.

3.2.3 Post Processing

Post-processing operations are performed so to map feature space to another space such that it is more suitable for the classifier. Post-processing operations involve normalization, dimensionality reduction, velocity and acceleration coefficients, etc.

Velocity and Acceleration Coefficients

To capture the *transitional* or *dynamic* information of the speech signal, velocity, and acceleration coefficients are appended with the static feature coefficients. The velocity (Δ) and acceleration ($\Delta\Delta$) features are extracted through the first and second derivative coefficients averaged over the interval, for the cepstrum time function. It is given by :

$$\Delta = \frac{dC}{dt}, \tag{3.2}$$

$$\Delta\Delta = \frac{d^2C}{dt^2}, \tag{3.3}$$

where C are the cepstrum coefficients. Although this triples the feature dimension, however, this increase in dimensions is more than compensated by the improved feature representation and hence, the performance of the replay SSD system.

3.3 Details of Dataset Used

3.3.1 Replay SSD

ASVspoof 2017 version 2.0 dataset

In chapter-4, experiments are performed on ASVspoof 2017 version 2.0 database. This dataset is based on the RedDots database [53], having a collection of genuine and spoof utterances. This speech data is collected from 177 replay sessions, and 61 unique replay configurations [53]. A replay configuration is a unique combination of a recording environment, playback device, and recording device. The database contains the recording of 42 speakers. Spoofed utterances were generated by replaying the genuine utterance in different environments and using different playback devices.

The ASVspoof 2017 Version 2.0 corpus consists of a training, development, and evaluation set. The first two subsets are to be used to develop the spoofing countermeasures. Metadata includes the file ID, speech ground truth label, phrase ID, and replay configuration details. The evaluation set is used to test the trained model and hence, evaluate the performance of trained models.

Table 3.1: Details About the ASVspoof 2017 Version 2.0 Database. After [52].

Datasets	**# Speakers**	**# Utternaces**	
		Genuine	**Spoof**
Training	10	1507	1507
Development	8	760	950
Evaluation	24	1298	12008
Total	42	3565	14465

ASVspoof 2019 PA

In chapter 4, ASVspoof 2019 challenge dataset is utilized for the developing the CM systems for the ASV systems [91]. The source signals for performing simulation are from the VCTK corpus [102]. This is the first database, which consists of all the three spoofing attacks SS, VC, and replay attacks, which are separated into two groups logical access and physical access scenarios. The first part of ASVspoof 2019 LA consists of SS and VC utterances, generated through the deep learning algorithms. The second part explores the replay attacks, which are generated using the simulation of the room acoustics under varying source/receiver positions through the approach reported in [91] and based upon an image-source method. This dataset consists of 107 speaker, 28890 genuine utterances and 189540 replay utterances in 27 various acoustic environments. Furthermore, the utterances are captured under various parameter settings for acoustic configurations given in [91].

ReMASC

In chapter-4, Realistic Replay Attack Microphone Array Speech Corpus (ReMASC) has been used, which is specifically designed to develop CM against replay spoofing attack. ReMASC dataset consists of four device comprising ~29000 utterances, among which ~25500 utterances are used for experiments as they are recorded by voice assistants [31]. Data collection is done in four acoustic environments, namely, outdoor environment (Env-A) containing random background noises, moving vehicle environment (Env-D), and two sets of indoor environments (Env-B and Env-C). Env-B is emulated in a silent meeting room with different device placements, while Env-C contains voice commands with background music player and TV sounds.

The statistics of the dataset used for performing experiments is shown in Table 3.2. Data partition is disjoint in terms of speakers. The data partition is done such that the proportion of environment-wise utterances in all the subsets remains the same as in the main dataset.

Table 3.2: Design of ReMASC Database. After [31].

Datasets	**Training**	**Development**	**Evaluation**
Genuine	5698	633	2118
Spoof	15458	1717	14162
Total	21156	2350	16280

VSDC

The VSDC dataset is designed for developing CM systems for VAs for multi-point replay attacks [6]. Further, chapter 4 performs cross-database evaluation between VSDC, ASVspoof 2017 version 2.0, and ASVspoof 2019 PA. This dataset consists

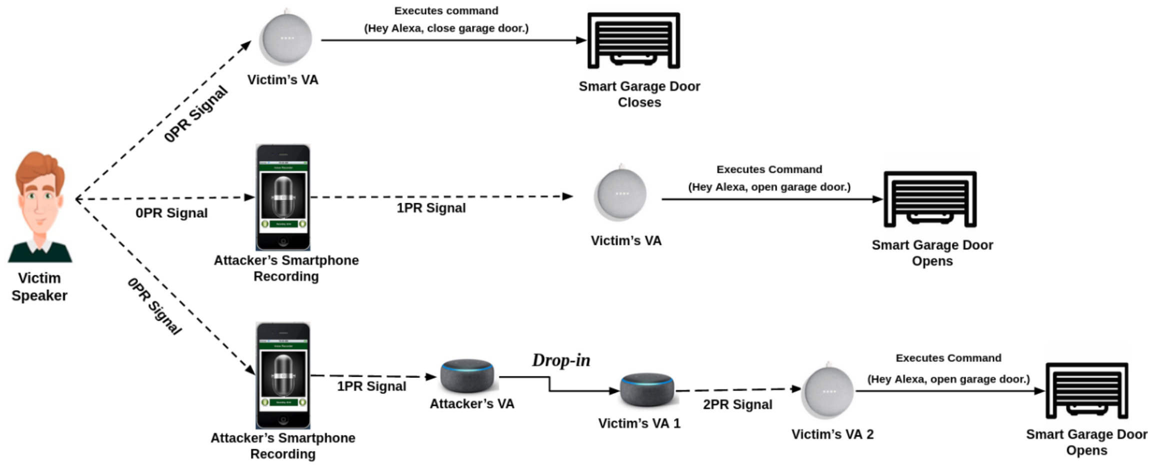

Figure 3.1: Schematic Illustrations and Realistic Scenarios of 0PR, 1PR, and 2PR. After [87].

of three types of utterances, namely, 0PR, 1PR, and 2PR. Figure 3.1 describes a scenario, which explains 0PR, 1PR, and 2PR speech. This scenario depicts a situation in which the victim is assumed to be someone close to the attacker, so that the attacker is able to record the required voice commands. This can be done by directing a normal conversation and using social engineering techniques. The speech of the victim, which is directly given as voice command to his/her VA, is genuine, and it is represented as 0PR. The attacker then discretely records the conversation and later executes 1PR and 2PR attacks, which are described in detail as follows.

- **One-Point Replay (1PR):** The use of smart devices is on the rise. These devices, such as baby monitors and smart doors, are equipped with built-in-microphones, loudspeakers/playback mechanism, and remote service access. These smart devices can be operated by VAs using remote service access. If the attacker has recordings of the victim's voice commands, he/she can playback the recorded speech in front of the victim's VA through his/her smartphone. The playback signal produced by the smartphone is called as 1PR utterance (as shown by the 1PR signal in Figure 3.1). Hence, the smartphone acts as the *first-point of replay attack* also known as 1PR attack.

- **Two-Point Replay (2PR):** The *Drop-in* feature initiates the two-way conversation between users using the VAs, similar to intercom devices [40]. This feature works between two different locations and different VAs owned by different users. This feature is enabled if the permission is allowed for the contact to *Drop-in*. Hence, the attacker can listen to the conversations at the location of the victim using *Drop-in* feature. Due to this feature, the attacker can record and playback victim's speech. Thus, we can establish that *Drop-in* feature can be used for prying into the possible personal information of the victim [1–3]. Furthermore, the attacker who has *Drop-in* permission can access the authorized speaker's VA and operate devices, such as smart bulb and smart doors. [40]. The scenario shown in Fig. 3.1, where the attacker is able to access the victim's 2^{nd} VA through the playback of 1PR utterance using the *Drop-in* feature (assuming the attacker has enabled the *Drop-in* feature permission at the victim's 1^{st} VA through social engineering). This playback of 1PR to the victim's 2^{nd} VA through the *Drop-in* feature of the 1^{st} VAs is known as Two-Point Replay (2PR) attack.

This dataset consists of 14050 utterances in total. Furthermore, there are 42 different command phrases of VAs used, which are recorded at a sampling frequency of 96 kHz. Furthermore, the dataset distribution is provided in Table 3.3.

Table 3.3: Design of VSDC Corpus Database. After [6].

Datasets	**Training**	**Evaluation**
Genuine (0PR)	3198	1371
1PR	3298	1427
2PR	3298	1397
Total	9794	4195

3.3.2 Dysarthria

Universal Access Speech Corpus (UA-Speech Corpus)

In chapter 4, and chapter 5 the Universal Access Speech Corpus (UA-Speech Corpus) is used for the development of the classification system based on severity-level dysarthria. The UA-Speech corpus [51] consists of dysarthric speech utterances from 19 subjects (15 males and 4 females). UA-Speech corpus is the largest database currently used for the dysarthric ASR, speech enhancement, and severity-level classification. The range of age for the speaker varies from 18-58 years. Furthermore, speakers with diverse speech intelligibility were taken into consideration. The intelligibility of each speaker was rated by naive human listeners on a scale of 0-100 %. The recording is done using a 8-channel microphone array with a sampling frequency of $16kHz$. For each speaker, data was collected in 3 recording sessions defined as *blocks*. These blocks consist of 255 words for recording, with 155 repeated words recording in all blocks and 100 distinct words in each block. The speech utterances are divided into the following categories:

- 10 English digits (0-9),
- 26 radio alphabets ("Alpha", "Bravo", "Charlie", etc.),

- 19 computer commands (“open”, “enter”, “delete”, etc.),
- 100 most common words from Brown corpus of written English),
- 100 uncommon words (naturalization, exploit, etc.) chosen from children’s novels.

Hence, 765 utterances are recorded in total, where each speaker records 255 utterances.

3.4 Details of Classifier Used

Once the feature representation is obtained, it is used to train an appropriate classifier. Given a speech sample X, the extracted feature is utilized for the binary classification problems, such as replay SSD. However, for severity-level classification of dysarthric speech, multi-class classification is required. Hence, various classifiers are implemented in this thesis work are discussed in the next sub-Section.

3.4.1 Gaussian Mixture Model (GMM)

A GMM back-end classifier was employed for genuine and spoofed speech detection [10]. It is the weighted sum of several multivariate Gaussian components. Gaussian mixture of N mixture components is represented as [10]:

$$P(y/G) = \sum_{j=1}^{N} w_j p(y/\mu_j, \Sigma_j), \tag{3.4}$$

where y is a D-dimensional feature vector, w_j is the weight corresponding to the multivariate Gaussian component $p(y/\mu_j, \Sigma_j)$ such that $\Sigma w_j = 1$. Here μ_j and Σ_j are $D \times 1$ mean vector and $D \times D$ dimensional covariance matrix, respectively. Thus, $p(y/\mu_j, \Sigma_j)$ is given by [10]:

$$p(y/\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{D/2}\Sigma_i^{1/2}}\, exp\Big\{ -\frac{1}{2}(y-\mu_i)^T \Sigma_j^{-1}(y-\mu_j)\Big\}. \tag{3.5}$$

The GMMs were trained using the expectation maximization (EM) criterion to obtain the maximum likelihood estimate (MLE) with random initialization. Individual GMMs trained on genuine and spoof training samples. The Log-Likelihood

Ratio (LLR) score is computed as [23].

$$LLR = log\frac{P(y|G_0)}{P(y|G_1)}, \tag{3.6}$$

where $P(y|G_0)$ and $P(y|G_1)$ are likelihood scores of genuine (G_0) and spoof (G_1), respectively.

3.4.2 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) are deep learning algorithms which uses the convolution operation in the architecture for processing the data. This convolution is done between the multidimensional input and multidimensional filter weight, known as *kernel*. The convolution operators are followed by the pooling-layer and non-linear activation operation. The combination of these three operation comprises a convolution layer, through which the features are extracted from input data. The *Fully-Connected* layers of perceptron are present in the CNN model for the classification. Further, CNN model extracts the feature similar to the human brain by using the convolutional layers and activation functions. CNN has been widely utilized for the image classification and pattern recognition. Hence, in this study to capture the energy-based features, the CNN model has been implemented.

Convolution Operation

In CNN, convolution operations are processed by sliding the kernel through the input matrix and processing the data. The kernel size is smaller than the input matrix. The convolution operator is represented as [56]:

$$G[g_h, g_w] = (F * K)[g_h, g_w] = \sum_{i=1}^{k_h} \sum_{j=1}^{k_w} K[i, j].F[g_h - i, g_w - j], \tag{3.7}$$

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

Padding Operation

The output matrix obtained by the convolution operation has lower dimension *w.r.t* the input matrix. Hence, for deeper convolution networks, the output will diminish. Furthermore, by applying the convolution operation, it can be observed that the effects of the boundary elements are less in comparison to the elements placed at the center, which is disadvantageous if the prominent features are present at the boundaries. Hence, to overcome these disadvantages, the input matrix is padded with random values (generally zeros). Hence, the dimension of the input matrix is increased and assist in capturing the information from the boundary elements. The padding size p, for a kernel size, $k \times k$ is calculated as [56]:

$$p = \frac{k-1}{2}. \tag{3.8}$$

Stride Convolution

In the convolution operation, the kernel overlaps each element in the input matrix. However, in the larger input matrix, this represents the computational inefficiency, because the calculations are done multiple times on every element of matrix, which consumes time and memory. Further, the capturing of the global feature and local are effected through the stride values. Additionally, the stride convolution contributes in dimensionality reduction resulting in a fewer calculations, which is desirable in many cases. The output dimension, n_{out} of convolution operation implemented by padding and striding, is estimated as [56]:

$$n_{out} = \frac{n_{in} + 2.p - k}{s} + 1, \tag{3.9}$$

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

Activation Layers

In neural network, the output of each smallest computation unit, namely, perceptron is passed through an activation function, which introduces the non-linearities in the neural network models. Hence, the output is differentiable, which assist in the back propagation and optimization of the weights. Activation function makes the deep neural networks suitable for the complex tasks, and generalized and adaptable to the data. The activation function makes the decision of enabling the perceptron in the next layer. For the activation function $\sigma(.)$, the output z for an input, x is defined as [81]:

$$z = \sigma(w.x + b), \tag{3.10}$$

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

Pooling Layer

Pooling layer is utilized for the dimensionality reduction without any significant reduction in the information present. The convolution layer output is generally provided as input to the pooling layer, through which the computational complexity of the CNN reduced, making the model faster to operate. The pooling layer captures the important features and make the network less susceptible to spatial movement from its kernel size. Therefore, the pooling layer do not affect the model performance, however, it increases the efficiency of the model.

Architecture Details

- **Replay SSD:** In this study which consists of four convolutional blocks, namely, Convolution 1, Convolution 2, Convolution 3, and Convolution 4, where each convolution block consists of 2-D convolutional layer, an activation layer, namely, Rectified Linear Activation (ReLU), and a max-pool layer. The input feature is of size $D \times 400$, where D represents the dimension of the feature vector. Convolution 1 and 4, each consist of a convolutional layer of kernel size of 3×3 with stride of 1 and padding of 2. Convolution 2 and 3 consist of convolutional layer of kernel size 5×5 with stride of 1 and padding of 2. Each of these convolutional layers is followed by ReLU and max-pool layer of kernel size 2×2. Lastly, the output of Convolution 4 is fed to Fully-Connected (FC1) linear layer with different hidden units. The model is trained using the Stochastic Gradient Descent (SGD) with a batch size of 32 and a learning rate of 0.001. Binary cross-entropy is selected for the loss calculation. Further, 10000 iterations are considered for training the model, which results in 106 number of epochs.

- **Dysarthria Severity-Level Classification:** In this study, CNN model was trained using Stochastic Gradient Descent (SGD) algorithm and 3 convolutional blocks each with kernel size 5×5, and 1 Fully-Connected (FC) layer [56]. The input feature is made of uniform size of $D \times 300$, where D is the dimension of the feature vector. Learning rate of 0.001 and cross-entropy loss is selected for loss estimation.

3.4.3 Light Convolutional Neural Network (LCNN)

Light-CNN (LCNN), a modified version of the neural network, has performed exceptionally for SSD task [95]. In LCNN, the non-linear activation functions are replaced with the Max-Feature-Map (MFM) activation layer, which is briefly discussed next.

Max-Feature-Map (MFM) Activation

MFM is a modified max-out function, which produces better generalization for distinct data distribution by learning with a small number of parameters. The MFM function is defined as [95]:

$$y_{ij}^{k} = max(x_{ij}^{k}, x_{ij}^{k+\frac{N}{2}}), \tag{3.11}$$

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

Architecture Details

- **Dysarthria Severity-Level Classification:** In this study, we utilized seven convolutional layers having MFM activation function followed by two-fully connected layers. The 1^{st} convolutional layer uses the kernel size of 5×5 and stride of, 1×1 and the following convolutional layer has a kernel size of 3×3 and stride of 2×2 with learning rate of 0.001. Weights of the LCNN are initialized using Xavier weight initialization technique [29].

3.4.4 Residual Neural Network (ResNet)

The vanishing gradient problem in CNN introduced a new classifier, namely, ResNet, which includes the skip connections into the architecture [21].

Skip Connection

The skip connections are implemented to resolve the vanishing gradient problem of deep neural networks. The vanishing gradient occurs in several layered neural networks [21]. The gradient estimation using the back propagation is usually less than 1, which provides mode stability to the model. However, in the large networks the gradient value is very small for the initial layers, which makes the effect of initial layer insignificant. Hence, the skip layer is utilized where it passes over one or more layers in neural network layers. This provides the gradient to flow during the back propagation, such that the initial layer gradient is not 0. Further, skip connections also enable the latter layers to learn information from the initial layers. The skip connections are of two types, namely, addition and concatenation. In the addition mode, the skip connection is added to the output from the layer of the network in an elementwise manner. In concatenation mode, the output is concatenated with the skip connection and used in the densely-connected networks. This forms the residual block of the ResNet model.

Residual Blocks

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

Architecture Details

- **Dysarthria Severity-Level Classification:** In this study we, have utilized 12 residual blocks out of which 9 are regular residual and 3 are downsampling residual blocks. The convolution layer of 5 with stride 2 is applied along with max pool layer of 2×2. The downsampling blocks are utilized to reduce the dimensionality of the feature maps. In the end, 1 fully connected is utilized for the multi-class classification. Similar to CNN and LCNN model, SGD with a batch size of 32 and a learning rate of 0.001 with 200 epochs.

3.5 Performance Measures

The performance of various feature sets was compared against the baseline feature set using various performance evaluation metrics, such as % classification Accuracy, Equal Error Rate (EER), Confusion Matrix, $F1-$ Score, J-Measure, Mathew's Correlation Coefficient (MCC), Jaccard's Index, Hamming Loss, and Linear Discriminant Analysis (LDA).

3.5.1 Confusion Matrix

Confusion matrix provides the overall performance of the classification models. It consists of $l \times l$, where l is the number of classes the data need to be classified. Confusion matrix arranges the prediction into following categories:

- **True Positive (TP)** :These are the samples which belong to a certain class and are correctly predicted.
- **True Negative (TN)**: These are the samples which do not belong to a certain class and are not predicted to that class but to any other class.
- **False Positive (FP)**: These are sample that do not belong to a certain class but are predicted to belong to that class.
- **False Negative (FN)**: These are the samples that belong to a certain class but are predicted to belong to any other class.

Confusion matrix is useful in evaluating the precision, recall, F1-score, J-Measure, Jaccard index, Hamming loss, and Mathews' Correlation Coefficient (MCC) of classification models.

3.5.2 % Classification Accuracy

Accuracy is the most simplified and powerful performance metrics used for the performance evaluation of deep learning models. % Classification Accuracy metric provides a fair evaluation for the balanced datasets. The accuracy (in %) is defined as:

$$Accuracy = \frac{\text{Number of correct prediction}}{\text{Number of total prediction}} \times 100\ \%. \tag{3.12}$$

3.5.3 Equal Error Rate (EER)

The EER is derived from the detection error trade-off (DET) curve, which represents the performance on detection tasks that involve the trade-off of error types [64]. In binary classification task, there are two types of errors, namely, false alarm rate ($P_{fa}(s)$) and miss rate ($P_{miss}(s)$). For arbitrary threshold s, these error rates are defined as [64]:

$$P_{fa}(s) = \frac{\text{number of class 1 trials with score} > s}{\text{total number of class 1 trials}}, \tag{3.13}$$

$$P_{miss}(s) = \frac{\text{number of class 2 trials with score} \leq s}{\text{total number of class 2 trials}}. \tag{3.14}$$

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

$$EER(\%) = P_{fa}(s_{EER}) = P_{miss}(s_{EER}). \tag{3.15}$$

3.5.4 $F1-$Score

It is a widely used statistical parameter for evaluating the performance of a model. It is estimated as the harmonic mean of the model's precision and recall [27]. In particular:

$$F1 - score = \frac{2TP}{2TP + FP + FN}. \tag{3.16}$$

The F1-score value ranges from 0 to 1, with a score closer to 1 signifying better performance.

3.5.5 J-Measure

J-statistic also known as Youden's J-statistic captures the performance of a dichotomous diagnostic test. It ranges between -1 and 1, where -1 indicates no agreement and +1 indicated full agreement between observation and prediction. Youden's J-statistic is given by [7]:

$$J - statistic = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1. \tag{3.17}$$

3.5.6 Mathews' Correlation Coefficient (MCC)

It shows the degree of association between the expected and actual class [65]. It is usually considered a balanced measure when comparing models. MCC is in the range of -1 to 1. It is given as [65]:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TN + FN)(TP_F N)(TN + FP)}}. \tag{3.18}$$

3.5.7 Jaccard Index

The Jaccard index is a measure of how similar and dissimilar the two classes are. Its value is in between 0 and 1. It is given in [11]:

$$JaccardIndex = \frac{TP}{TP + FP + FN}. \tag{3.19}$$

3.5.8 Hamming Loss

It takes into account incorrectly predicted class labels. All the classes and test data are normalized for prediction error (prediction of an inaccurate label) and missing error (prediction of a relevant label). Hamming loss can be calculated as [22]:

$$\text{Hamming Loss} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j), \tag{3.20}$$

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

3.5.9 Linear Discriminant Analysis (LDA)

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

3.6 Chapter Summary

In this chapter, the details of the experimental setup for the replay SSD, and severity-level classification. Additionally, the speech processing methodologies, classifiers used, and the performance measures were also discussed. In the next chapter, we discuss the Teager Energy-based Teager energy cepstral coefficients features.

CHAPTER 4

TECC Feature Set

4.1 Introduction

In this chapter, brief description of Teager Energy Cepstral Coefficients (TECC) is presented for the development of the CM system for replay SSD, and severity-level classification of dysarthria. Furthermore, the cross-database evaluation is performed between the VSDC, ASVspoof 2017, and ASVspoof 2019 using the TECC feature set with CQCC as baseline feature set. Further, TECC is extended toward the multichannel audio input by using the relative change in the energy between the microphone arrays on ReMASC dataset. Finally, the TECC is extended severity-level classification of dysarthric speech.

4.2 Teager Energy Operator (TEO)

In the production and perception of the speech signal, energy established a connecting link. The energy is transmitted through the sound medium through the vibration mechanism depicted through the mass-spring simple harmonics motion (SHM). For a mass-spring system, the energy is proportional to the squared product of amplitude and frequency [33]. In particular,

$$E = \frac{1}{2} m A^2 \omega^2, \tag{4.1}$$

$$E \propto A^2 \omega^2, \tag{4.2}$$

where m is the mass of the suspension, A is the amplitude of oscillation, and ω is the frequency of oscillation. This analogy, as in [69], has been used to model amplitude modulated frequency modulated (AM-FM) signal with the help of an energy operator called Teager Energy operator (TEO) [85]. Teager's work on non-linear modelling of the human speech production in [86] are used to model, and detect modulations in speech resonances.

TEO captures the running estimates of the signal energy, which can be represented for a continuous-time domain signal as [85]:

$$\Psi\{x(t)\} = \left(\frac{dx}{dt}\right)^2 - \frac{d^2}{dt^2}x(t). \tag{4.3}$$

Furthermore, the TEO for the discrete-time signal is estimated as [85]:

$$\Psi\{x[n]\} = x_n^2 - x_{n-1}x_{n+1} = A^2 sin^2\omega \approx A^2\omega^2, \tag{4.4}$$

where x_n be a discrete-time signal expressed as $x_n = Acos[\omega n + \theta]$, the $x_{n-1} = Acos[\omega(n-1) + \theta]$, and $x_{n+1} = Acos[\omega(n+1) + \theta]$. Additionally, the TEO captures the reverberation present in the signal and also suppresses the additive noise, which is discussed in the next Section.

4.3 Properties of TEO

4.3.1 Capturing of Reverberation

A speech signal, especially in a closed space, such as a small room, gets reflected multiple times due to infinite transmissions and reflections from surfaces, such as wall, furniture, and even people. However, this is not limited to closed spaces, and can even occur in open spaces, such as in a forest [47]. The formation of these multiple transmissions and reflections is called as *reverberation*. These reflections tend to decay with distance and time as they are absorbed by the surrounding surfaces in the space. The majority of reverberation energy is found in reflections of the 1^{st} order (i.e., with only one deviation from the original path) and 2^{nd} order (i.e., with two deviations). These reverberation effects of 1^{st} order and 2^{nd} order reverberation can be observed from the utterances of VSDC, shown in Fig. 4.1. The Panel-II of Fig. 4.1 shows the Teager energy profiles of genuine (i.e., 0PR), 1PR, and 2PR speech utterances. It can be observed that the 0PR has substantially less effect of reverberation as compared to the 1PR and 2PR utterances. Hence, the capturing the reverberation acts as the discriminating acoustic cues for the replay SSD.

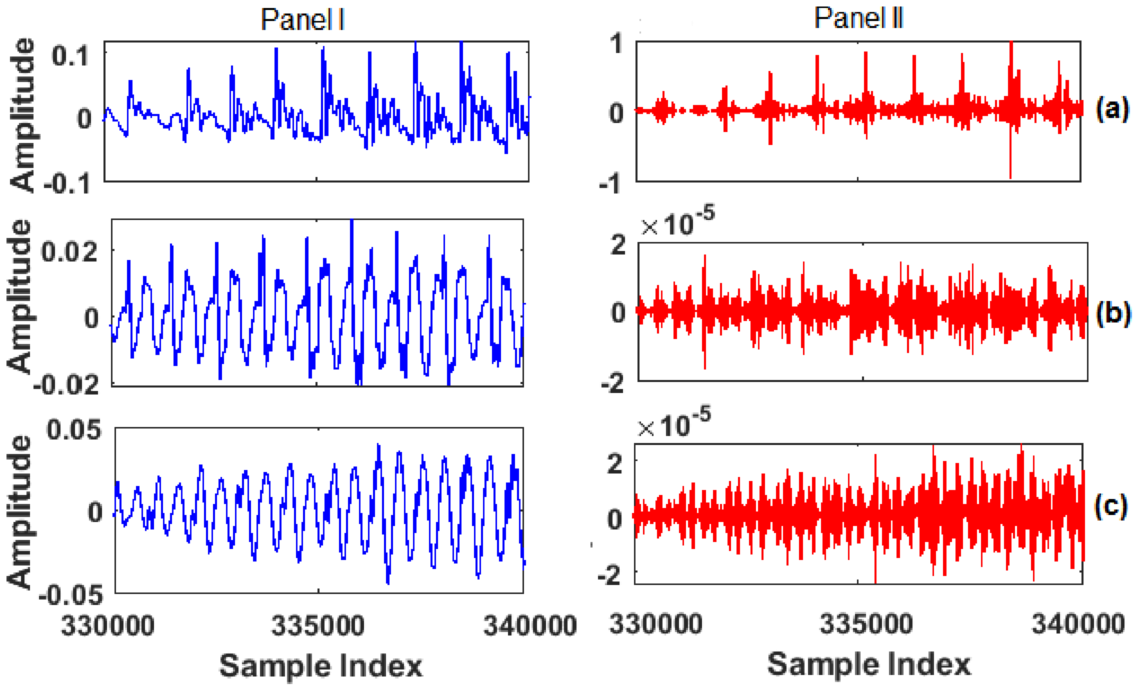

Figure 4.1: Time-domain Signal (Panel I) and its corresponding Teager Energy Profiles (Panel II) for the Three Types of Utterances: (a) 0PR, (b) 1PR, and (c) 2PR. After [87].

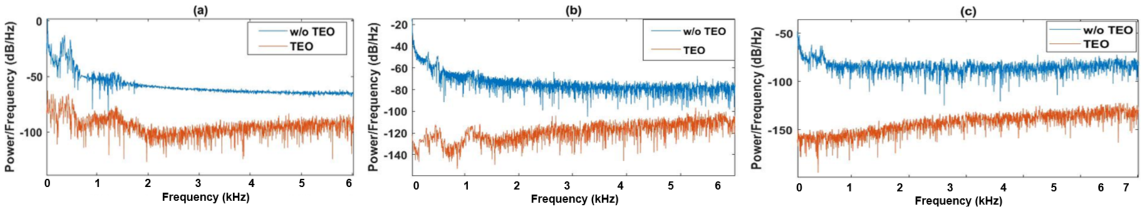

Figure 4.2: The Power Spectral Density (PSD) of the (a) 0PR, (b) 1PR, and (c) 2PR Speech Segment. After [87].

4.3.2 Noise Suppression by TEO

Let us consider a speech signal $x(n)$ corrupted with various types of additive noises, such as Gaussian, Poisson, and uniform, so that the resultant output signals for 1PR and 2PR are defined as [37], [38]:

$$y_1[n] = x[n] + \eta_1[n], \tag{4.5}$$

$$y_2[n] = x[n] + \eta_1[n] + \eta_2[n]. \tag{4.6}$$

where $\eta_1[n]$ and $\eta_1[n] + \eta_2[n]$ represents the additive noise present in 1PR and 2PR scenarios, respectively. The TEO of these 1PR and 2PR resultant signal is calculated as:

$$\Psi\{y_1[n]\} = \Psi\{x[n]\} + \Psi\{\eta_1[n]\} + 2 * \tilde{\Psi}\{x[n], \eta_1[n]\}, \tag{4.7}$$

$$\begin{aligned}\Psi\{y_2[n]\} = \Psi\{x[n]\} + \Psi\{\eta_1[n]\} + \Psi\{\eta_2[n]\} + 2 * \tilde{\Psi}\{x[n], \eta_1[n] + \eta_2[n]\} + \\ 2 * \tilde{\Psi}\{\eta_1[n], \eta_2[n]\},\end{aligned} \tag{4.8}$$

$$\eta[n] = \eta_1[n] + \eta_2[n], \tag{4.9}$$

where $\tilde{\Psi}$ represents the cross-Teager energy operator (CTEO) between the two signals $a[n]$, and $b[n]$ and it is defined as [12]:

$$\begin{aligned}\tilde{\Psi}\{a[n], b[n]\} = a[n]b[n] - 0.5.[a[n+1]b[n-1] \\ +a[n-1]b[n+1]].\end{aligned} \tag{4.10}$$

Moreover, in eq. (4.7) and eq. (4.8) the $x[n]$, $\eta_1[n]$, $\eta_2[n]$, and $\eta[n]$ are zero-mean and statistically-independent signals. Therefore, the expected value of $\tilde{\Psi}\{x[n], \eta[n]\}$, $\tilde{\Psi}\{x[n], \eta[n]\}$ is zero. Further the $E[\Psi\{\eta_1[n]\}] = E[\Psi\{\eta_2[n]\}] \approx 0$. Hence, the expected value of resultant TEO equation is defined as [60]:

$$E[\Psi\{y_1(n)\}] \approx E[\Psi\{x[n]\}]. \tag{4.11}$$

$$E[\Psi\{y_2(n)\}] \approx E[\Psi\{x[n]\}]. \tag{4.12}$$

Fig. 4.2 shows the Power Spectral Density (PSD) of TEO (orange line) and without TEO (blue line) of 0PR, 1PR, and 2PR for the same speech segment as shown in Fig. 4.2. The noise suppression capabilities of the TEO can be observed through the power difference between the TEO and without TEO segment. The power difference between PSD of TEO of signal and without TEO signal decreases as the frequency increases for the 0PR, 1PR, and 2PR. Hence, the key idea for using the TECC feature set is to exploit the noise suppression capabilities of TEO for the classification of the genuine *vs.* replay utterance.

4.4 TECC Feature Extraction

The functional block diagram of the TECC feature set is shown in Fig. 4.3 [49]. Given that speech is a multicomponent signal and that the hearing is a process of detecting energy in subbands [45], practically it should be decomposed into several subbands. Therefore, the speech signal is first bandpass filtered using a linearly-spaced Gabor filterbank, resulting in N narrowband (subband) filtered signals. Due to the Heisenberg's uncertainty principle in signal processing framework, the Gabor filter is known to have *optimal* time-frequency resolution [61]. Moreover, it has been found that the linearly-spaced frequency bins enable good resolution both in lower and higher frequency regions of characteristics desired to capture attributes of replay spoof. Furthermore, in order to estimate the energy of each of the narrowband filtered signals, TEO is used to estimate energy with high temporal resolution, utilizing only *three* consecutive samples for energy estimation. To that effect, the application of TEO to each of the filtered narrowband signal results in N TEO profiles. These TEO profiles are passed through frame-blocking, and averaged with a short window of 20 ms and with a window shift of 10 ms, followed by logarithmic operation to compress the data. Lastly, Discrete Cosine Transform (DCT) is applied for feature decorrelation and energy compaction along with Cepstral Mean Normalization (CMN) to derive TECC feature set [47].

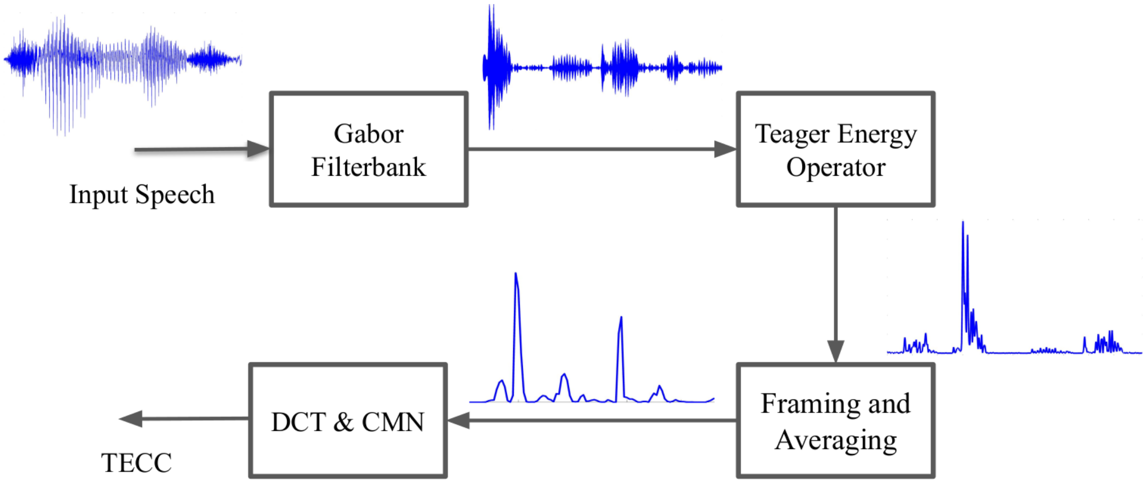

Figure 4.3: Functional block diagram of TECC feature extraction. After [87].

4.5 Cross-Teager Energy Cepstral Coefficients (CTECC)

4.5.1 Analysis of Cross-Teager Energy Operator (CTEO)

In sub-Section 4.2, TEO is implemented for single channel analysis. Hence, to track the cross-energies between two channels, CTEO is developed in [46], and can be denoted as $\Psi_{cr}[\cdot]$. CTEO is a nonlinear quadratic operator, which estimates the relative rate of change of energies between signals. The Cross-Teager Energy (CTE) between the two *real-valued* signals, $x(t)$ and $y(t)$ in continuous-time-domain is represented as [57]:

$$\Psi_{cr}\{x(t), y(t)\} = (\dot{x}(t)\dot{y}(t)) - (x(t)\ddot{y}(t)), \tag{4.13}$$

$$\Psi_{cr}\{y(t), x(t)\} = (\dot{y}(t)\dot{x}(t)) - (y(t)\ddot{x}(t)). \tag{4.14}$$

From eq. (4.13) and eq. (4.14), the non-commutative property of CTEO is observed, i.e., $\Psi_{cr}[x(t), y(t)] \neq \Psi_{cr}[y(t), x(t)]$ [46], [12]. Using eq. (4.13), the *average CTEO* ($\Psi_{cr}^{avg}[\cdot]$) between the continuous-time *real-valued* signals is estimated as [12]:

$$\Psi_{cr}^{avg}\{x(t), y(t)\} = \frac{1}{2}(\Psi_{cr}\{x(t), y(t)\} + \Psi_{cr}\{y(t), x(t)\}). \tag{4.15}$$

However, the definition of CTEO can be extended to complex-valued signals as given in [19]. Furthermore, for the discrete-time signals $x(n)$ and $y(n)$, average cross-Teager energies are estimated as:

$$\begin{aligned}\Psi_{cr}^{avg}\{x(n), y(n)\} = x(n)y(n) - 0.5[&x(n+1)y(n-1) \\ &+x(n-1)y(n+1)].\end{aligned} \tag{4.16}$$

From eq. (4.16), the excellent time resolution of the CTEO can be observed. Subsequently, the later part of the thesis deals with the real-valued continuous-time-domain representation of speech signal, which can be further extended in discrete-time.

Let us consider the signal $x_i(t)$ in N-sensor microphone array, where $i \in [1, N]$ and $x_i(t)$ is represented as:

$$x_i(t) = s_i(t) + n_i(t), i = 1, 2, ..., N, \tag{4.17}$$

where $s_i(t)$ and $n_i(t)$ represent the original speech signal and additive noise in i^{th} the sensor, respectively. The additive noise component is assumed to be zero-mean and Wide Sense Stationary (WSS) Gaussian random process. The impact of reverberation is neglected because acoustical reverberation are spectrally similar and cannot be separated from the natural speech.

The output signal of each sensor $x_i(t)$ is decomposed using a suitable filterbank into L subband signals, and subband filtered signal is represented as:

$$x_{i_j}(t) = x_i(t) * g_j(t), \quad j = 1, 2, ..., L, \tag{4.18}$$

where '*' represents the convolution and $x_{i_j}(t)$ represents the subband filtered signal obtained for the i^{th} channel and j^{th} subband filter in the filterbank. Considering two sensor input (p, q) and j^{th} subband filter of the filterbank, the CTE will be expressed as:

$$\Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\} = (\dot{x}_{p_j}(t)\dot{x}_{q_j}(t)) - (x_{p_j}(t)\ddot{x}_{p_j}(t)). \tag{4.19}$$

From the eq. (4.4), eq. (4.17), and eq. (4.19), we obtain:

$$\begin{aligned}\Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\} = \Psi_{cr}\{s_j(t)\} + \Psi_{cr}\{n_{p_j}(t), n_{q_j}(t)\} \\ + \Psi_{cr}\{s_j(t), n_{q_j}(t)\} + \Psi_{cr}\{n_{p_j}(t), s_j(t)\}.\end{aligned} \tag{4.20}$$

The replay noise is represented by the last three terms on the Right-Hand Side (RHS) of eq. (4.20). Taking expectation operator ($E[\cdot]$) on eq. (4.20), we get:

$$E[\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)\}] = E\{\Psi_{cr}[s_j(t)]\} + E\{\Psi_{cr}[n_{p_j}(t), n_{q_j}(t)]\}. \tag{4.21}$$

The last two terms of RHS side in eq.(4.20) are zero-mean and hence, the expectation operator is zero [57]. However, the second term represents the error in eq. (4.21) [62]. Hence, the modified equation is given as:

$$E\{\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)]\} = E\{\Psi_{cr}[s_j(t)]\} + error. \tag{4.22}$$

Let us denote τ the concentration of noise power within the subband filter's passband. Using Cauchy–Schwartz inequality for two random variables X and Y, we have [9]:

$$|E(XY)|^2 \leq E(X^2)E(Y^2), \tag{4.23}$$

where *(XY)* is the inner product between the random variables X and Y. Therefore, using eq. (4.23), the relation between the noise power, we obtain:

$$|\tau_{(pq)_j}| \leq \tau_{p_j}\tau_{q_j}, \tag{4.24}$$

where τ_{p_j} is the noise power concentration of the j^{th} subband and p^{th} channel. Moreover, τ_{p_j} is proportional to the *error* term in eq. (4.22), where the *error* term is the varying whereas the source signal through the bandpass filter remains the same throughout the analysis. For ASR application, the desirable speech signal representation should contain the least amount of noise component. Hence, the representation with minimum *error* is chosen for ASR application as explained in [77]. Whereas, for replay SSD, it is necessary to emphasize the distorted channel information and hence, we have chosen the channels, which corresponds to maximum *error* in eq. (4.22).

By maximizing the *error*, the additional acoustical representation can be obtained. With respect to the analysis of CTEO, we have ${}^{N}C_2$ possibilities of channel-pairs for each i^{th} subband. Estimating average CTE for all channel-pairs and then choosing the one with the highest average energy is a feasible, however, computationally expensive approach. To reduce the computational complexity, the two channels with the highest average Teager energy can be chosen and CTE between those two channels can be utilized, represented by the first block of mean and max in Fig. 4.4. Furthermore, among the set of the one average CTE and two Teager energies, the subband filtered signal with the maximum energy is selected for classification between genuine and replay utterances, namely, Maximum Energy Signal (MES) represented by second block of mean and max in Fig. 4.4. Mathematically, MES can be represented as:

$$\begin{aligned} MES =& max_{(p,q)}(E\{\Psi_{cr}^{avg}[x_{p_j}(t), x_{q_j}(t)]\}, E\{\Psi_{cr}[x_{p_j}(t)]\}, \\ & E\{\Psi_{cr}[x_{q_j}(t)]\}). \end{aligned} \tag{4.25}$$

From eq. (4.25), the MES contains the maximum distortions, such as acoustical environment and intermediate device responses, these non-linearities are captured by the MES. Hence, for the replay SSD, the MES is selected for further processing. However, for the severity-level classification of dysarthria the linguistic information has greater significance *w.r.t.* the environmental acoustic cues. Hence, among the set of the one average CTE and two Teager energies, the subband filtered signal with the minimum energy is selected for severity-level classification, namely, Minimum Energy Signal (MiES), represented as :

$$\begin{aligned} MiES =& min_{(p,q)}(E\{\Psi_{cr}^{avg}[x_{p_j}(t), x_{q_j}(t)]\}, E\{\Psi_{cr}[x_{p_j}(t)]\}, \\ & E\{\Psi_{cr}[x_{q_j}(t)]\}). \end{aligned} \tag{4.26}$$

4.5.2 CTECC Feature Extraction Procedure

Fig. 4.4 and Fig. 4.5 shows the functional block diagram of the CTECC for the development of CM system for replay attacks and classification of dysarthria based on severity, respectively. Input speech signal is recorded using different microphone arrays with different sampling frequencies. Hence, the input speech of each channel is re-sampled at 16 kHz sampling frequency. Each of the signal from N-channel microphone array is processed through a Gabor filterbank, which possess excellent time-frequency resolution (because the Fourier transform of a Gaussian function is also a Gaussian). Further, Gaussian belongs to the class of infinitely differentiable functions, in particular, $\mathbb{C}^{\infty}$ and hence, they have faster decay in frequency-domain [61]). The Gabor filterbank consist of linearly-spaced 160 and 40 subband filters for replay SSD and severity-level classification of dysarthria, respectively. Further, TEO profile for each subband filtered signal is obtained. Furthermore, average of the TEO_{nj} are compared, where $n \in [1, N]$ and $j \in [1, 160]$ for replay SSD and $j \in [1, 40]$ for dysarthric severity-level classification. Further, from the discussion in sub-Section 4.5.1, the two channels p and q are selected such that they have maximum average TEO which provides the maximum environmental acoustic cues. However, for the dysarthric classification based on severity-level, the two channels p and q are selected such that they have minimum average TEO, which contains maximum linguistic information. Further, using eq. (4.16) on the p and q, the average CTEO is estimated. Windowing is performed on the subband filtered signal with window size of 25 ms and window shift of 10 ms, which provides m frames. Averaging on each frame is performed, which provides the average energy for a frame in consideration. Then logarithm operation is performed, which is followed by Discrete Cosine Transform (DCT) to obtain the cepstral representation. The input feature for the classifier is obtained by the concatenation of the static, Δ, and ΔΔ components, described in Table 4.1.

Table 4.1: Dimension of CTECC Feature for Various Application. After [89].

Application	**Subband Filter**	**Static**	**Δ**	**ΔΔ**
Replay SSD	40	40	40	40
Severity-level Classification	160	70	70	70

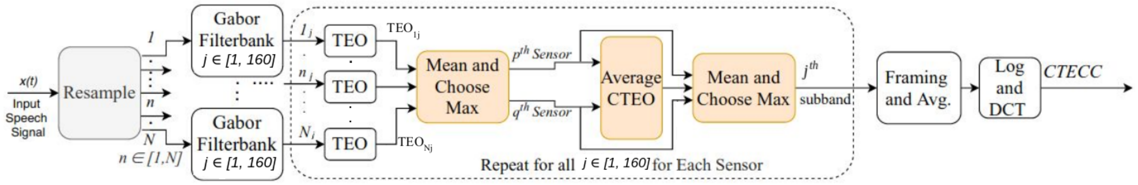

Figure 4.4: Functional Block Diagram of CTECC Feature Extraction for Replay SSD. After [89].

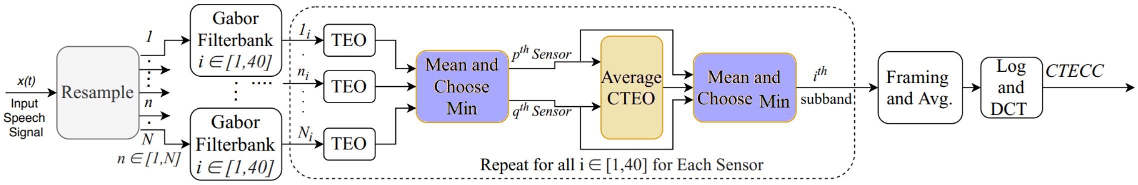

Figure 4.5: Functional Block Diagram of CTECC Feature Extraction for Severity-Level Classification of Dysarthria. After [89].

4.6 Experimental Results of TECC on Cross-Database Evaluation

In this Section, we present the experimental results for various cases of cross-dataset evaluation. We trained our model for three datasets, results of which are explained next:

4.6.1 Training on Complete VSDC Dataset

For this case, we utilized the complete VSDC dataset for training. Evaluation was performed on ASVSpoof 2017 v2.0 dataset and ASVSpoof 2019 PA dataset. The experimental results for these cases are described in this sub-Section.

Testing on ASVSpoof 2017 v2.0

- **With GMM classifier:** From Table 4.2, it can be observed that the TECC (linear) feature set achieves an EER of 27.63% on the evaluation set of ASVSpoof 2017 dataset and outperforms the rest of the feature sets. In particular, TECC feature set achieves a relative improvement of 38.21% EER on the evaluation set *w.r.t.* the baseline CQCC feature set.

- **With CNN classifier:** From Table 4.2, it can be observed that the TECC (linear) achieves EER of 24.07% and 26.19% on the development and evaluation sets, respectively.

Testing on ASVSpoof 2019 PA

- **With GMM classifier:** From Table 4.5, it can be observed that TECC variants achieve better performance as compared to the CQCC, MFCC, and LFCC. Furthermore, out of the three variants of TECC, TECC (inv-Mel) achieves relatively the best performance of 44.11% EER. One of the potential reasons for high EERs in the case of testing on ASVSpoof 2019 PA dataset is that VSDC consists of *real* replay signals, whereas the replay utterances in ASVSpoof 2019 PA dataset are *simulated*.

Table 4.2: Results (in % EER) training on VSDC and testing on ASVSpoof 2017. After [87].

Classifier	Training Dataset	VSDC	
	Testing Dataset	ASVSpoof 2017	
		Dev.	Eval.
GMM	CQCC	31.74	44.72
	MFCC	39.17	35.88
	LFCC	30.41	37.27
	TECC (Linear)	**20.01**	**27.63**
	TECC (Mel)	28.87	34.84
	TECC (Inv-Mel)	33.14	36.59
CNN	TECC (Linear)	24.07	26.19
	TECC (Mel)	28.68	33.05
	TECC (Inv-Mel)	33.15	33.51

4.6.2 Training on ASVSpoof 2017 Training Dataset

For this case, we trained our model on the training set of the ASVSpoof 2017 v.20 dataset. For evaluation, binary classification was performed for two cases: 1) 0PR and 1PR as genuine and spoof class, respectively. 2) 0PR and 2PR as genuine and spoof class, respectively.

Evaluation on VSDC with 0PR and 1PR as Genuine and Spoof (i.e., 0-1PR)

- **With GMM classifier:** From Table 4.3, it should be noted that minimum EER of 31.65% is achieved using TECC (Mel) leading to a relative improvement of 39.27% in ERR as compared to the CQCC baseline. It should also be noted that the remaining two variants of TECC, i.e., TECC (linear), and TECC (inv-Mel) also performed reasonably well as compared to the CQCC, MFCC, and LFCC.

- **With CNN classifier:** When CNN is used as the classifier, it can be observed that out of the three variants of TECC, TECC (linear) achieves relatively the best performance of 34.60% EER. Further, TECC (linear) gave the best performance on CNN. This was also the case when the training dataset was VSDC and the testing dataset was ASVSpoof 2017 v2.0.

Table 4.3: Results (in % EER) training on ASVSpoof 2017 and testing on VSDC. After [87].

Classifier	Training Dataset	ASVSpoof 2017	
	Testing Dataset	VSDC (0PR, 1PR)	VSDC (0PR, 2PR)
GMM	CQCC	52.12	43.16
	MFCC	45.28	46.87
	LFCC	47.78	49.13
	TECC (Linear)	41.29	40.88
	TECC (Mel)	**31.65**	31.96
	TECC (Inv-Mel)	38.04	34.45
CNN	TECC (Linear)	34.6	**22.97**
	TECC (Mel)	43.83	27.27
	TECC (Inv-Mel)	43.39	36.61

Evaluation on VSDC with 0PR and 2PR as Genuine and Spoof (i.e., 0-2PR)

- **With GMM classifier:** From Table 4.3, TECC (Mel) achieves relatively the best performance of 31.96% EER, with relative improvement of 25.94% in EER w.r.t. CQCC baseline system.

- **With CNN classifier:** Amongst the three variants of TECC, TECC (linear) achieves relatively the best performance of 22.97% with CNN. However, the remaining variants, i.e., TECC (Mel) and TECC (linear) also perform relatively better than the rest of the feature sets.

4.6.3 Training on ASVSpoof 2019 Training Dataset

Evaluation on VSDC with 0PR and 1PR as Genuine and Spoof (i.e., 0-1PR)

- **With GMM classifier:** From Table 4.4, it can be observed that TECC and its variants outperform the CQCC, MFCC, and LFCC feature sets. In particular, TECC (Mel) achieves relatively the best performance of 33.45% EER, which leads to percentage improvement of 26.33% in EER *w.r.t.* the CQCC baseline.

Evaluation on VSDC with 0PR and 2PR as Genuine and Spoof (i.e., 0-2PR)

- **With GMM classifier:** From Table 4.4, it is noted that TECC (inv-Mel) achieves relatively the best performance of 34% EER, which leads to percentage improvement of 17.31% in EER w.r.t. the CQCC baseline.

Table 4.4: Results (in % EER) training on ASVSpoof 2019 PA and testing on VSDC dataset using GMM classifier.After [87].

Training Dataset	ASVSpoof 2019 PA	
Testing Dataset	VSDC	
	0PR-1PR	0PR-2PR
CQCC	45.41	41.12
MFCC	47.03	38.59
LFCC	41.88	41.52
TECC (Linear)	41.65	39.97
TECC (Mel)	**33.45**	40.60
TECC (Inv-Mel)	35.07	**34.00**

Table 4.5: Results (in % EER) training on VSDC and testing on ASVSpoof 2019 PA dataset using GMM Classifier. After [87].

Training Dataset	VSDC	
Testing Dataset	ASVSpoof 2019 PA	
	Dev.	Eval.
CQCC	48.66	49.46
MFCC	49.97	49.80
LFCC	49.70	49.45
TECC (Linear)	47.34	48.17
TECC (Mel)	46.38	47.59
TECC (Inv-Mel)	**42.96**	**44.11**

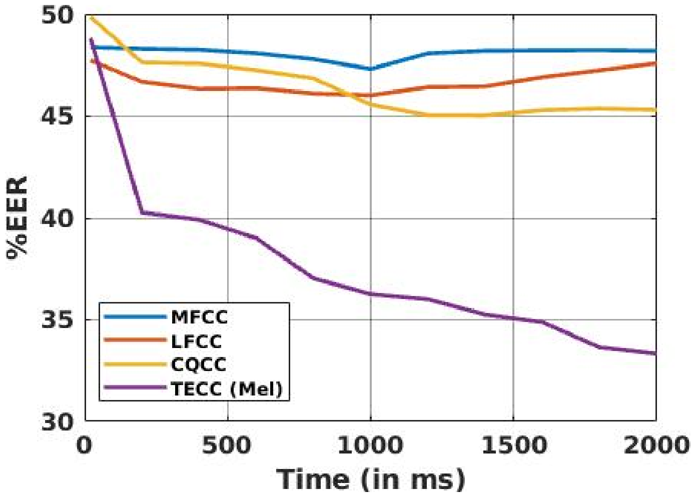

Figure 4.6: Latency period *vs.* % EER between the MFCC, LFCC, CQCC, and TECC (Mel) feature sets. After [87].

4.6.4 Analysis of Latency Period

In this study, we have also investigated the latency period for TECC feature set *w.r.t* the other feature sets considered in this study. Latency period, represents the performance evaluation *w.r.t* different durations of speech segment in an utterance and estimating the % EER. The utterance duration ranges from 20 ms to 2 seconds, with an interval of 200 ms. Further, the utterance duration is selected by considering the number of frames. Fig. 5.4 shows comparison between the MFCC, LFCC, CQCC, and TECC (Mel) for training on ASVSpoof 2017 and testing on VSDC 0PR-1PR. It can be observed that the TECC feature set outperformed the state-of-the-art features for the cross-database evaluation. Moreover, it can be observed that TECC feature set gave reduced % EER in short duration of speech utterance compared to the other feature sets. Furthermore, the % EER converges to the minimum value as the speech duration provided to the model of SSD system increases. This is due to the fact that more information is provided to the CM system for the classification and hence, the decreases in % EER is observed. Additionally, the performance of feature set is better if for a low latency period, the performance is high, which indicates the faster classification by the model and thus, indicating suitability of a SSD system from the perspective of practical deployment.

4.7 Results using CTECC

In this study, experiments are performed using the proposed CTECC feature set with GMM as a classifier and compared against the performance of the deep learning-based approach utilized in [32]. The proposed CTECC-GMM architecture and the deep learning-based architecture reported in [32], both exploits the multi-channel information representation for replay SSD, and the performance comparison for these two approaches is shown in Table 4.6 w.r.t. the number of channels utilized in each device. The architecture in [32] was considered as baseline architecture in our work. It can be observed that the proposed CTECC feature set performs relatively better than the baseline architecture for devices D1, D2, and D4, when all the available channels in the microphone array were utilized for feature representation. Whereas, the comparable performance is observed for the device D3. Furthermore, the results obtained using proposed CTECC feature set are compared against the other state-of-the-art feature sets, such as MFCC, CQCC, LFCC, and TECC. The results are shown in Table 4.7. It can be observed that the proposed CTECC feature set performs better than the other feature sets, except for device D1. On the whole, the proposed CTECC feature set is useful representation for the replay SSD for VAs, where multi-channel information can be exploited.

Table 4.6: Results (in % EER) on evaluation set for the proposed CTECC-GMM architecture *vs.* architecture proposed in [32] for various devices.

Device	**Channels Utilized for Replay SSD**													
	1		**2**		**3**		**4**		**5**		**6**		**7**	
	[32]	**CTECC**	**[32]**	**CTECC**	**[32]**	**CTECC**	**[32]**	**CTECC**	**[32]**	**CTECC**	**[32]**	**CTECC**	**[32]**	**CTECC**
D1	16.6	22.0	**14.9**	**8.84**	-	-	-	-	-	-	-	-	-	-
D2	23.7	25.80	19.5	15.74	16.7	16.33	**15.4**	**13.01**	-	-	-	-	-	-
D3	23.7	24.63	19.1	17.31	17.6	19.755	17.0	19.10	17.1	19.71	**16.5**	**16.53**	-	-
D4	27.5	29.79	21.5	21.47	20.6	21.19	21.3	20.3	20.7	20.25	19.9	21.15	**19.8**	**16.41**

Table 4.7: Results (in % EER) on development and evaluation set w.r.t. various feature sets and for various devices. After [89].

Device	**D1**		**D2**		**D3**		**D4**	
Feature Set	**Dev**	**Eval**	**Dev**	**Eval**	**Dev**	**Eval**	**Dev**	**Eval**
MFCC	9.16	7.98	16.87	26.02	19.71	20.37	12.23	26.99
CQCC	2.88	11.9	4.59	28.68	4.07	23.91	1.88	29.392
LFCC	2.26	8.04	3.45	20.09	4.75	19.32	3.43	23.18
TECC	9.15	22.0	12.34	25.80	10.85	24.63	13.72	29.79
CTECC	**0.86**	**8.84**	**1.25**	**13.01**	**0.87**	**16.53**	**0.67**	**16.41**

Table 4.8: Performance Measure for classification experiments for genuine and replay utterances. After [89].

	Devices Used for Replay SSD							
	D1		**D2**		**D3**		**D4**	
	Dev	**Eval**	**Dev**	**Eval**	**Dev**	**Eval**	**Dev**	**Eval**
Feature	**F-Measure**							
MFCC	0.9245	0.9165	0.9070	0.8139	0.8889	0.8738	0.9160	0.7998
CQCC	0.9705	0.9110	0.9492	0.7763	0.9578	0.8057	0.9677	0.7355
LFCC	0.9646	0.9386	0.9613	0.8396	0.9605	0.8719	0.9727	0.8090
TECC	0.9746	**0.9449**	0.9425	**0.8988**	0.9575	**0.9339**	0.9327	**0.8813**
CTECC	**0.9843**	0.9438	**0.9827**	0.8980	**0.9851**	0.8710	**0.9941**	0.7975
J-measure								
MFCC	0.8358	0.7843	0.7862	0.5143	0.7031	0.6008	0.7767	0.4918
CQCC	0.9368	0.7418	0.8907	0.4534	0.9038	0.4917	0.9260	0.4261
LFCC	0.9258	0.8297	0.9065	0.5557	0.8984	0.6055	0.9317	0.5178
TECC	0.8812	0.5577	0.8496	0.5618	0.8571	0.5199	0.8107	0.5105
CTECC	**0.9690**	**0.8304**	**0.9629**	**0.7286**	**0.9657**	**0.6412**	**0.9847**	**0.9103**

4.7.1 Analysis of Latency Period

In this study, we have also considered the analysis of latency period, where speech segment of different durations ranging from 20 ms to 2 seconds with an interval of 200 ms is presented to the model and the % EER is calculated. The utterance duration is selected by considering the number of frames. Fig. 5.4 shows comparison between the CTECC, MFCC, CQCC, LFCC, and TECC feature sets *w.r.t* latency period. It can be observed that the CTECC provides the lowest % EER *w.r.t* the other single channel feature sets. Further, the CTECC considers the multi-channel input, which assist in capturing the cues of acoustical environment. However, the response of CM system (in terms of % EER) converges as the speech duration increases. This is due to the fact that more information is provided to the CM system for the classification and hence, the decrease in % EER is observed.

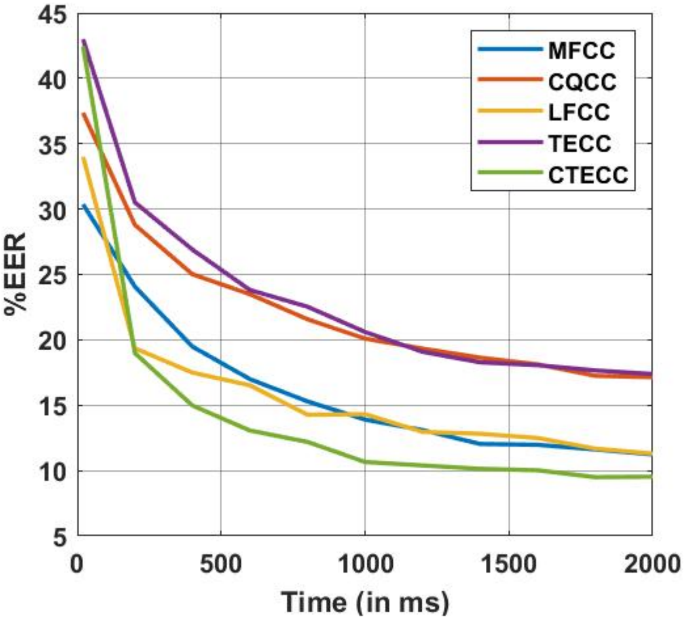

Figure 4.7: Latency period *vs.* % EER comparison between MFCC, CQCC, LFCC, TECC, and CTECC. After [89].

4.8 Results on Severity-Level Classification

4.8.1 Results using TECC

The results obtained in % classification accuracy using various features sets and classifiers are reported in Table 5.1. It can be observed that the TECC performs relatively better than the baseline MFCC feature set with classification accuracy of 97.18%, 94.63% and 98.02% (i.e., absolute improvement of 1.98 %, 1.41 %, and 1.69 %) for CNN, LCNN, and ResNet classifiers, respectively. Furthermore, it was also observed that there was decrease in % classification accuracy by varying parameters in CNN model. This might be due to overfitting of the model. Furthermore, TECC performs the better than then baseline MFCC feature set for CNN, LCNN, and ResNet classifiers explored in [41]. Moreover, it was observed that optimum results of TECC were obtained on linear scale. The analysis provided in Section

4.3 along with experimental results obtained using various classifiers shows that the TECC can be the best possible choice for the severity-level classification of dysarthric speech.

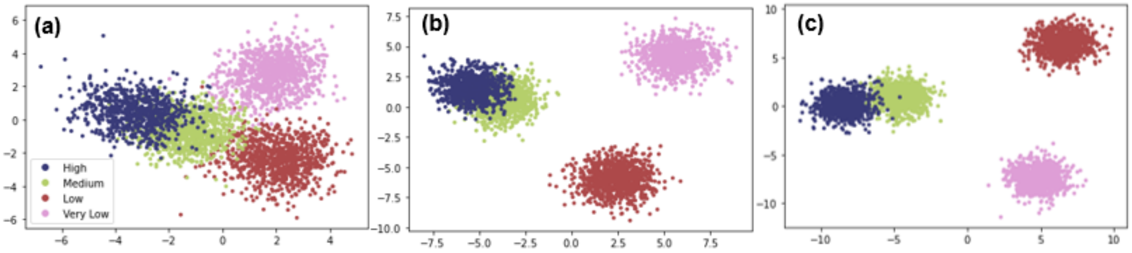

Figure 4.8: Scatter plots obtained using LDA for (a) MFCC, (b) LFCC, and (c) TECC. After [43]. Best viewed in colour.

Table 4.9: Results (in % Classification Accuracy) for various classification systems. TECC → Linear frequency scale used. After [43].

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

Table 4.10: Results (in % Accuracy) between deep learning classifiers.

Feature Set	**% Classification Accuracy**			
	SVM	**CNN**	**LCNN**	**ResNet**
MFCC	75.70	95.20	93.22	96.33
LFCC	77.40	96.32	94.07	97.17
TECC	**79.37**	**97.12**	**94.63**	**98.02**

As mentioned in [54], the cepstral features perform better on noisy signal. In [101], the noise in dysarthric speech increases with increase in severity-levels. Hence, experiments were also performed on the spectral features *w.r.t* proposed and baseline features with all the three classifiers. It was observed that the cepstral features gave remarkably better % classification accuracy on all the classifiers. Further, the CNN classifier is compared with the traditional SVM classifier in one-versus-one approach, shown in Table 4.10. Hence, it can be inferred that more the severity-level, more is the speech production noise.

Furthermore, Table 4.11 shows the confusion matrices for the TECC, MFCC, and LFCC for ResNet model. It can be observed that TECC reduces the misclassification errors, especially for high severity-level dysarthria, and overall performance of the TECC is relatively better than the MFCC, and LFCC. Furthermore, $F1$-score, MCC, Jaccard index, and Hamming loss are estimated for all the cepstral features as shown in Table 4.12. It can be observed from Table 4.12 that the TECC feature set outperforms the other cepstral features for all the evaluation metrics, indicating relatively better feature discriminative power of TECC.

Table 4.11: Confusion matrix obtained for MFCC, LFCC, and TECC using ResNet. After [43].

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

Table 4.12: Various statistical measures for MFCC, LFCC, and TECC. After [43].

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

4.8.2 Analysis of Latency Period

We analysed latency period for TECC, LFCC, and MFCC feature sets as shown in Figure 5.4. The latency period of the trained model is estimated by computing the % classification accuracy *w.r.t.* varying durations of test speech segment in a test utterance. For analysis of latency period, we chose the duration of the utterances varying from 100 ms to 3000 ms. The better performing model w.r.t. latency period should produce the larger accuracy for short speech segments. Moreover, it can be observed that the TECC gave significant % classification accuracy in a limited duration speech utterance of $<$ 500 ms. On the contrary, MFCC and LFCC shows increment in accuracy after a relatively longer utterance duration of 1000 ms. Hence, these results signifies the suitability of TECC for practical dysarthric speech classification system deployment.

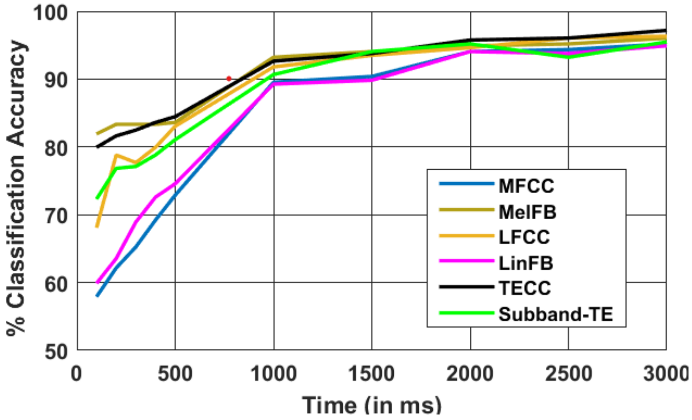

Figure 4.9: Latency period *vs.* % classification accuracy comparison between MFCC, MelFB, LFCC, LinFB, TECC, and Subband-TE. After [88]. Best viewed in colour.

4.8.3 Results using CTECC

The % classification accuracy of baseline STFT and CTECC on CNN is shown in Table 4.13. It can be observed that the CTECC (min) performs better with classification accuracy of 95.76% than the baseline STFT and CTECC (max) on CNN model. Furthermore, the performance analysis shown in the Table 4.14 using statistical parameters, such as $F1-$Score, MCC, Jaccard Index, and Hamming Loss, also shows that the CTECC (min) shows better linguist information capturing capabilities from the dysarthric speech compared to then CTECC (max) and STFT feature set on CNN model. In addition to it, Table 5.2 shows the confusion matrix of STFT, CTECC (max), and CTECC (min) feature set. It can be observed from the Table 4.12 that the false prediction is reduced by the CTECC (min) in comparison to baseline STFT and CTECC (max) feature set, which all the more supports the fact that CTECC (min) is capable of capturing the linguistic or relevant discriminant information better than STFT and CTECC (max) feature set.

Table 4.13: % Classification Accuracy for Baseline STFT and CTECC Feature Set. After [88].

Feature Set	**CNN**
Spectrogram	91.72
CTECC_max	91.24
CTECC_min	**95.76**

Table 4.14: Performance Evaluation for Various Feature Set. After [88].

Feature Set	F1-Score	MCC	Jaccard Index	Hamming Loss
STFT	0.87	0.83	0.776	0.124
CTECC_max	0.91	0.88	0.84	0.087
CTECC_min	**0.96**	**0.94**	**0.91**	**0.042**

Table 4.15: Confusion Matrix for STFT, and CTECC Feature Set. After [88].

Feature Set	Severity	High	Medium	Low	Very Low
STFT	**High**	63	6	3	3
	Medium	10	79	3	1
	Low	3	4	79	7
	Very Low	1	2	1	89
CTECC (Max)	**High**	62	10	2	1
	Medium	4	85	1	1
	Low	1	3	88	1
	Very Low	1	4	2	86
CTECC (Min)	**High**	**70**	3	2	0
	Medium	3	**90**	0	0
	Low	1	3	**87**	2
	Very Low	0	1	0	**92**

4.8.4 Analysis of Latency Period

Finally, we also analysed the latency period for CTECC (Min) and CTECC (Max) feature sets as shown in Figure 5.4. The latency period of the trained model is estimated by computing the % classification accuracy *w.r.t.* varying durations of test speech segment in a test utterance [80]. For latency period analysis, we chose the duration of the utterances varying from 20 ms to 400 ms. The better performing model w.r.t. latency period should produce the larger accuracy for short speech segments. Moreover, it can be observed that the CTECC gave significant % classification accuracy in a short duration of *w.r.t* CTECC_max. Hence, these results signifies the suitability of CTECC for deployment of practical dysarthric speech classification system.

4.9 Chapter Summary

In this chapter, we presented the significance of TECC feature set for development of CM system for replay SSD, and severity-level classification of dysarthria. Further, the potential application of TECC in the generalization of the CM system and severity-level classification of dysarthria has been supported through the performance measures, such as Hamming loss, J-measure, and latency period. Furthermore, CTECC, an extension of the TEO was utilized for the replay SSD for VAs

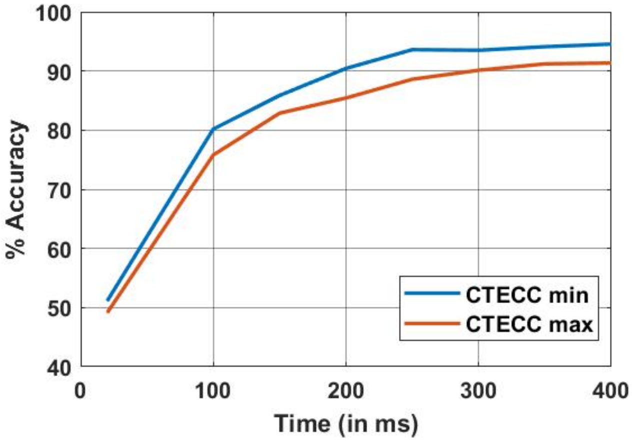

Figure 4.10: Latency Period *vs.* % Accuracy Comparison Between CTECC min and CTECC max. After [88].

and severity-level classification of dysarthria, where the impact of multi-channel microphone in capturing the relative change in energy was observed. Finally, the next chapter proposes the Squared Energy Cepstral Coefficients (SECC) for the severity-level classification of dysarthria.

CHAPTER 5
SECC Feature Set

5.1 Introduction

In this chapter, we discuss the Squared Energy Cepstral Coefficients (SECC), which is an extension of our work on TECC for dysarthria as discussed in chapter 4. The sub-Section 5.2 describe the feature extraction procedure of SECC. Further, sub-Section 5.3 discusses the Squared Energy Operator (SEO) and TEO profiles for the dysarthric speech. Next, the Section 5.4 and Section 5.5 presents the experimental results on the UA-corpus and the performance analysis obtained by the SECC, respectively. Finally, Section 5.6 summarizes the chapter.

5.2 Proposed Work

In the signal processing literature, the energy of the speech signal $x(t)$ is estimated by calculating the integral of square of absolute operation across the entire signal under consideration, i.e., estimating the squared L^2 norm of the signal, referred to as SEO [68]. This energy estimation method is based on linear filtering theory (specifically, Parseval's energy equivalence, the total energy of a signal, i.e., L^2 norm is conserved in the frequency-domain and this is also the condition of existence of inverse for several *linear* transforms, such as Fourier, Gabor (STFT), and wavelet transforms), which can only represent the linear components of the speech generation process [86].

For SECC extraction, these narrowband output signals from Gabor filterbank are squared to estimate corresponding energies. Next, these narrowband energies are segmented with similar number of frames and window overlap. Temporal averaging for each frame is estimated (i.e., L^2 norm of each subband signal) to get N-D *subband Squared Energy representation (subband-LE)*. Discrete Cosine Transform (DCT) is applied on *subband Squared energy representations* to obtain the SECC. The functional block diagram representation of TECC and SECC feature sets is shown

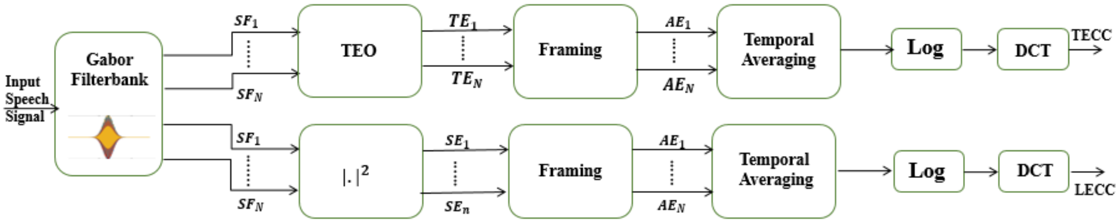

Figure 5.1: Functional block diagram of the proposed TECC and SECC feature sets. (SF: Subband filtered signal, SE: Squared linear energies, TE: Teager energies, AE: Averaged energies over frames). After [42].

in Fig 5.1. Throughout this paper, TECC and SECC features extracted using linear frequency scale and for both the feature sets, DCT does the job of feature decorrelation, energy compaction, and feature vector dimensionality reduction.

5.3 Squared Energy Operator Profile Analysis

Here, we analyse the TEO profiles around the 1^{st} formant frequency (i.e., $F_1 = 500Hz$) for the utterance *w.r.t.* the same text material for normal *vs.* severity-levels. Panel I of Fig. 5.2 shows the subband filtered signal around 1^{st} formant (F_1) frequency using a linearly-spaced Gabor filter, and Panel II shows corresponding TEO profiles. Fig. 5.2(a), Fig. 5.2(b), Fig. 5.2(c), Fig. 5.2(d), and Fig. 5.2(e) shows the analysis for normal, very low, low, medium, and high severity-levels, respectively. It can be observed that the TEO profile for normal speech shows *bumps* within two consecutive Glottal Closure Instants (GCIs), which are known to indicate non-linearities in speech production mechanism [74]. Furthermore, it can also be observed that the quasi-periodicity in glottal excitation source decreases with increase in severity-level (as observed via aperiodic TEO profile) indicating *disruption* in the rhythmic quasi-periodic movements of vocal folds due to dysarthria. Moreover, it is all the more significant in high severity dysarthric condition. Furthermore, as the severity-level increases, the neuro-motor impairment also increase, which leads to increased disruption in vocal fold closure and loosing *structural* periodicity. From Panel III of Fig. 5.2, which shows the SEO profiles around 1^{st} formant frequency for vowel /e/, it can be observed that the SEO is capable of maintaining the periodicity in the speech wave produced by dysarthric speaker, which are not captured by TEO due to possible decrease in non-linearities. Hence, it can be said that as the dysarthric severity-level increases, the linearities in speech signal increases.

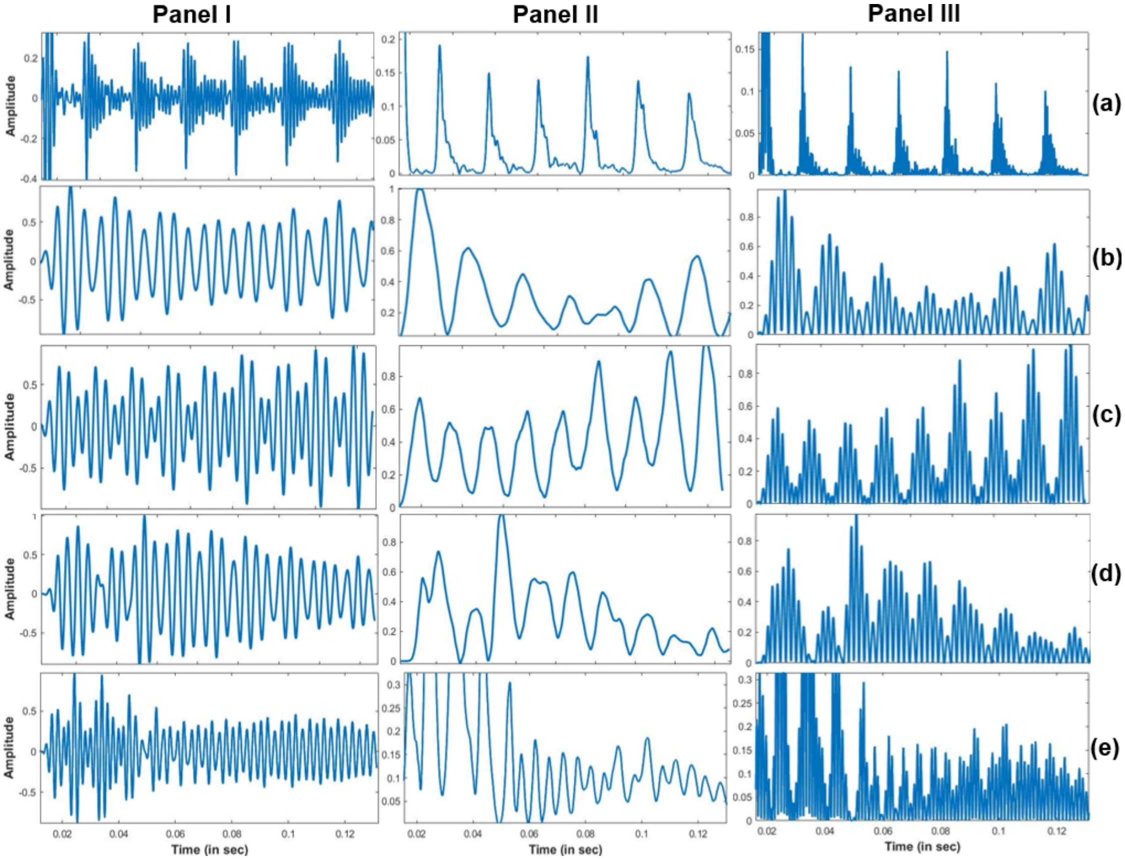

Figure 5.2: Subband filtered signal (for vowel $/e/$) for male speakers around $1^{st} formant F_1 = 500Hz$ (Panel I), corresponding TEO profile (Panel II), and corresponding $|.|^2$ envelope (Panel III) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. After [42].

5.4 Experimental Results

The results obtained as % classification accuracy using various feature sets are reported in Table 5.1. It can be observed that SECC performs relatively better than the baseline MFCC and TECC with classification accuracy of 1.7% (4.23% / 0.99%) than baseline MFCC and 0.1.41% (0.56% / 0.28%) than TECC on CNN (LCNN / ResNet) classifier systems, respectively. Furthermore, SECC performs better than the baseline MFCC explored in [41]. The analysis in the next Section, along with the % classification accuracy obtained using various classifiers, indicate that the linearities in speech production mechanism increases with increase in dysarthric severity-level.

% Please add the following required packages to your document preamble:

Table 5.1: Results (in % Classification Accuracy) For various Classification Systems. After [42].

Feature Set	% Classification Accuracy			
	CNN	LCNN	ResNet	SVM
MFCC	96.32	92.09	95.33	75.70
TECC	96.61	95.76	96.04	79.37
SECC	**98.02**	**96.32**	**96.32**	**88.31**

5.5 Performance Evaluation

Furthermore, Table 5.2 shows the confusion matrices for MFCC, TECC, and SECC for ResNet model. It can be observed that SECC reduces the misclassification errors corresponding to the different severity-levels, indicating the better performance of SECC *w.r.t.* TECC and MFCC. Furthermore, performance of SECC *w.r.t.* TECC and MFCC is also analysed using $F1$-Score, MCC, Jaccard Index, and Hamming Loss as shown in Table 5.3. It can be observed from Table 5.3 that SECC performs better than the TECC for the dysarthic severity-level classification.

Table 5.2: Confusion matrix for MFCC, TECC, and SECC using CNN. After [42].

Feature	Severity	High	Medium	Low	Very Low
MFCC	**High**	67	4	3	1
	Medium	2	90	0	0
	Low	1	1	91	0
	Very Low	1	0	0	92
TECC	**High**	72	1	2	0
	Medium	2	90	0	0
	Low	1	1	91	0
	Very Low	0	0	0	93
SECC	**High**	74	1	0	0
	Medium	2	90	0	0
	Low	1	0	92	0
	Very Low	0	0	0	93

Table 5.3: Various Statistical Measures of MFCC, TECC, and SECC. After [42].

Feature Set	F1-Score	MCC	Jaccard Index	Hamming Loss
MFCC	0.96	0.95	0.82	0.036
TECC	0.97	0.96	0.95	0.025
SECC	**0.98**	**0.97**	**0.96**	**0.019**

5.5.1 Linear Discriminate Analysis (LDA)

Capability of SECC to classify severity-level is also validated by LDA scatter plots, which projects the higher-dimensional feature space to the lower-dimension [36]. Here MFCC, TECC, and SECC features are projected to the 2-*D* space to get the scatter plots for various severity-levels of dysarthria. Fig. 5.3(a), Fig. 5.3(b), and Fig. 5.3(c) shows the LDA plots of MFCC, TECC, and SECC, respectively. From the Fig. 5.3, it can be observed that for SECC, the variance of each severity-level clusters is less resulting in relatively better performance of SECC, which increases the interclass distance between the clusters than the MFCC and TECC.

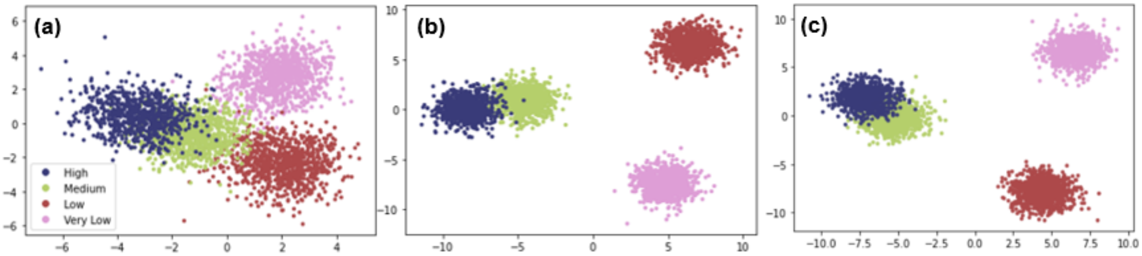

Figure 5.3: Scatter plots obtained using LDA for (a) MFCC, (b) TECC, and (c) SECC. After [42]. Best viewed in colour.

5.5.2 Latency Period Analysis

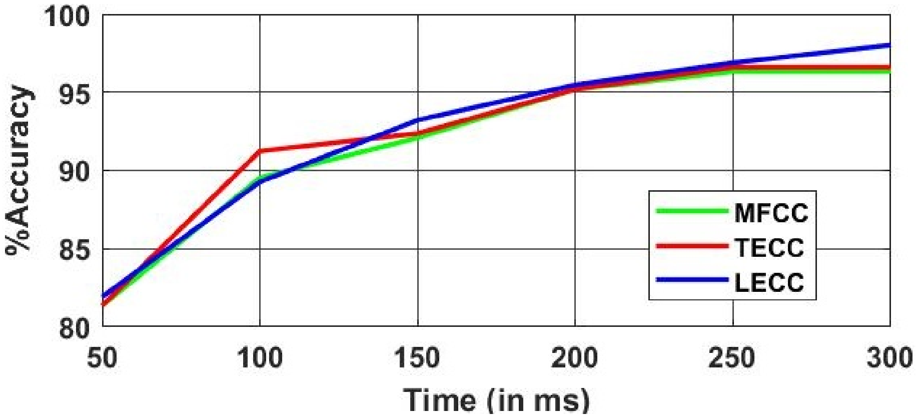

Figure 5.4: Latency period *vs.* % classification accuracy comparison between MFCC, TECC, and SECC. After [42]. Best viewed in colour.

Latency period for SECC *w.r.t.* TECC and MFCC were also analysed as shown in Fig. 5.4. The latency period was calculated using the % classification accuracy on varying test utterance. The utterance was varied from 50 ms to 300 ms. For short speech segments, the better performing model should produce higher accuracy in terms of latency period. From the Fig. 5.4, it can be observed that SECC gives consistent and relatively better classification accuracy in short duration of time as 60 ms. Furthermore, TECC and MFCC gives increased classification accuracy for speech segment of 100 ms and 250 ms, respectively. Hence, these results signifies the practical suitability of SECC in dysarthric speech severity-level classification.

5.6 Chapter Summary

This chapter presented the usefulness of the SECC feature set for the severity-level classification of dysarthria. Furthermore, we analysed the effect of the linear *vs.* non-linear energy operator for the analysis and classification of the severity-level of the dysarthric speech, indicating the presence of linearities in the dysarthric speech. Lastly, these hypotheses are tested and supported by performing the experiments using CNN, LCNN, and ResNet classifiers and the performance measures. Lastly, the next chapter concludes with the chapter summary and conclusion of this thesis.

Chapter 6
Summary and Conclusions

6.1 Summary of the Thesis

The work presented in the thesis aims at developing CM for replay SSD, and severity-level classifier for dysarthria using feature-based approach. The practical potential of TECC was observed from the generalization of the CM system through cross-database evaluation. Similarly, the TECC feature set provides the significant improvement in the classification of dysarthric speech-based on severity-level, which supports the hypothesis of TECC on capturing of source excitation information during speech production mechanism. In addition to TECC, CTECC was also introduced, which utilizes the multi-channel input for capturing the discriminative environmental acoustic cues for the replay SSD by selecting the optimum channel, which has the maximum relative energy. Similarly, for the severity-level classification of dysarthria, CTECC selects the optimum channel, which has the minimum relative energy (maximum linguistic energy). Finally, SECC feature set is implemented for the severity-level classification of dysarthria, through which the direct proportionality between presence of linearity and the severity-level of dysarthria was observed.

6.2 Conclusions

The following conclusions can be drawn from the thesis works:

- Significance of TECC feature set in developing the CM for replay detection and severity-level classification of dysarthria.
- The generalization of CM provides practical relevance for the deployment of the SSD systems.
- The testing database has greater significance in the cross-database evaluation.

- The concept of CTEO for selecting optimum channel for designing the countermeasure system for replay SSD and severity-level classification of dysarthria.
- The SEO-based feature aids in justifying the direct proportionality relation between severity-level of dysarthria and the presence of linearities.
- Various data augmentation techniques for dysarthria classification.
- Cross-database evaluation for dysarthria for generalization

6.3 Limitations of Thesis Work

- The generalization of the CM system by TECC feature set is dependent on the testing scenarios.
- Analysis of CTECC in the replay SSD and dysarthric severity-level classification.
- The presence of linear component in dysarthria is not generalized through cross-database evaluation.
- The motor control distortion between brain and primary speech mechanism or brain and secondary speech producing mechanism remains unknown.

6.4 Future Research Direction

- Explore CM model to generalize the SSD systems using various signal processing and deep learning methods.
- The significance of source-filter interaction feature for developing replay SSD and severity-level classification of dysarthria.
- Cross database evaluation of dysarthria for generalization of severity-level classification.
- Dysarthric speech enhancement for voice assistant and ASR systems.
- Automatic Speech Recognition models for Dysarthric patients.

CHAPTER 7
List of Publications from Thesis

(1) **Anand Therattil**, Priyanka Gupta, Piyushkumar K. Chodingala, Hemant A. Patil, "Cross-Database Evaluation for Detection of one point and two point replay attacks," **accepted** in: The Speaker and Language Recognition Workshop, Speaker Odyssey, Bejing, China, June 28-July 1 2022.

(2) Aastha Kachhi, **Anand Therattil**, Ankur T. Patil, Hardik B. Sailor, Hemant A. Patil "Dysarthric Speech Severity-Level Analysis and Classification Using Teager Energy Cepstral Features", **submitted** in: September INTERSPEECH Incheon Korea, 28-22 September 2022.

(3) Aastha Kachhi, **Anand Therattil**, Ankur T. Patil, Hardik B. Sailor, Hemant A. Patil "Analysis of Non-Linearities in Normal vs. Dysarthric Speech For Severity-Level Classification", **submitted** in: Signal Processing and Communications (SPCOM), IISc, Bangalore, 11-15 July 2022.

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

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

(6) Madhu R. Kamble, **Anand Therattil**, Hemant A. Patil, M. Ali Basha Shaik, and Vikram Vij "Impact of Acoustic Environment and Microphone Array for Voice Assistant Systems using Smoothed Teager Energy Features", **Rejected** in: The Speaker and Language Recognition Workshop (Speaker Odyssey) Bejing, China, June 28-July 1 2022,

(7) **Anand Therattil**, Ankur T. Patil and Hemant A. Patil "On Significance of Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants", **Rejected** in: The Speaker and Language Recognition Workshop (Speaker Odyssey), Bejing, China, June 28- July 1 2022.

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

References

[1] Alexa's drop in feature makes eavesdropping easy. https://www.wfmynews2.com/article/news/local/2-wants-to-know/alexas-drop-in-feature-makes-eavesdropping-easy/83-528254751 {Last Accessed: 26-02-2022}.

[2] Amazon disputes claims that echo show's drop-in feature is a security risk. https://techcrunch.com/2017/06/28/amazon-disputes-claims-that-echo-shows-drop-in-feature-is-a-security-risk/ {Last Accessed: 26-02-2022}.

[3] How to call another alexa device in a different house. https://robotpoweredhome.com/how-to-call-another-alexa-device-in-a-different-house/ {Last Accessed: 26-02-2022}.

[4] R. Acharya, H. Kotta, A. T. Patil, and H. A. Patil. Cross-Teager energy cepstral coefficients for replay spoof detection on voice assistants. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Ontario, Canada,pp. 6364–6368*, 6-11 June 2021.

[5] S. R. Atcherson, A. E. DeLaune, K. Hadden, R. I. Zraick, R. J. Kelly-Campbell, and C. P. Minaya. A computer-based readability analysis of consumer materials on the american speech-language-hearing association website. *Contemporary Issues in Communication Science and Disorders*, vol.41, pp.12–23, 2014.

[6] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik. Voice spoofing detection corpus for single and multi-order audio replays. *Computer Speech & Language*, vol.65:pp.101132, January 2021.

[7] V. Bewick, L. Cheek, and J. Ball. Statistics review 10: further nonparametric methods. *Critical care*, vol.8,pp.1–4, 2004.

[8] C. Bhat and H. Strik. Automatic assessment of sentence-level dysarthria intelligibility using blstm. *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, 4pp. 322–330, 2020.

[9] R. Bhatia and C. Davis. A Cauchy-Schwartz inequality for operators with applications. *Linear Algebra and Its Applications*, vol.223, pp.119–129, 1995.

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

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

[12] A.-O. Boudraa, J.-C. Cexus, and K. Abed-Meraim. Cross ψ b-energy operator-based signal detection. *The Journal of the Acoustical Society of America (JASA)*, vol.123,pp.4283–4289, 2008.

[13] H. Byeon. Comparing ensemble-based machine learning classifiers developed for distinguishing hypokinetic dysarthria from presbyphonia. *Applied Sciences*, vol. 11, pp. 2235, 2021.

[14] W. Cai, D. Cai, W. Liu, G. Li, and M. Li. Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion. In *INTERSPEECH, pp.17–21, Stockholm, Sweden*, 2017.

[15] R. Cardoso, I. Guimarães, H. Santos, R. Loureiro, J. Domingos, D. de Abreu, N. Gonçalves, S. Pinto, and J. Ferreira. Frenchay dysarthria assessment (fda-2) in parkinson's disease: cross-cultural adaptation and psychometric properties of the european portuguese version. *Journal of neurology*, vol. 264, pp. 21–31, 2017.

[16] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden voice commands. In *25th (USENIX) Security Symposium (USENIX) Security 16), pp.513–530, Austin, TX, USA*, 2016.

[17] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW), pp. 1–7, SAN FRANCISCO, CA*. IEEE, 2018.

[18] N. Carlini and D. Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, pp. 1-7, san francisco, ca. In *2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA*, 24-24 May 2018.

[19] J.-C. Cexus and A.-O. Boudraa. Link between cross-Wigner distribution and cross-Teager energy operator. *Electronics Letters*, vol. 40pp. 778–780, 2004.

[20] H. Chandrashekar, V. Karjigi, and N. Sreedevi. Spectro-temporal representation of speech for intelligibility assessment of dysarthria. *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 390–399, 2019.

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

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

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

[24] W. Diao, X. Liu, Z. Zhou, and K. Zhang. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices, pp. 63–74, Scottsdale Arizona USA*, 7 November 2014.

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

[26] G. Esposito and P. Venuti. Understanding early communication signals in autism: a study of the perception of infants' cry. *Journal of Intellectual Disability Research*, vol.54:pp.216–223, 2010.

[27] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, vol.27:pp.861–874, 2006.

[28] M. Fernández-Díaz and A. Gallardo-Antolín. An attention long short-term memory based system for automatic classification of speech intelligibility. *Engineering Applications of Artificial Intelligence*, vol. 96, pp. 103976, 2020.

[29] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and*

Statistics, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[30] Y. Gong and C. Poellabauer. Protecting voice controlled systems using sound source identification based on acoustic cues. In *2018 27th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9*. IEEE, 2018.

[31] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer. ReMASC: Realistic replay attack corpus for voice controlled systems. INTERSPEECH 2019,*pp. 2355–2359, Graz, Austria*, 2 Jul 2019.

[32] Y. Gong, J. Yang, and C. Poellabauer. Detecting replay attacks using multichannel audio: A neural network-based method. *IEEE Signal Processing Letters*, vol. 27, pp. 920–924, 2020.

[33] R. C. Guido. Enhancing Teager energy operator based on a novel and appealing concept: Signal mass. *Journal of the Franklin Institute*, vol. 356, pp. 2346–2352, 2019.

[34] T. Gunendradasan, B. Wickramasinghe, P. N. Le, E. Ambikairajah, and J. Epps. Detection of replay-spoofing attacks using frequency modulation features. In *INTERSPEECH, pp. 636–640636–640*, Hyderabad, India, Sept. 2018.

[35] N. Gurevich and S. L. Scamihorn. Speech-language pathologists' use of intelligibility measures in adults with dysarthria. *American journal of speech-language pathology*, vol. 26, pp. 873–892, 2017.

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

[37] F. Jabloun and A. E. Cetin. The Teager energy based feature parameters for robust speech recognition in car noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume vol. 1, pp. 273–276, Phoenix, AZ, USA, 1999.

[38] F. Jabloun, A. E. Cetin, and E. Erzin. Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Processing Letters*, vol. 6, pp. 259–261, 1999.

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

[40] A. Javed, K. M. Malik, A. Irtaza, and H. Malik. Towards protecting cyber-physical and Iot systems from single-and multi-order voice spoofing attacks. *Applied Acoustics*, vol. 183, pp. 108283, 2021.

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

[42] A. Kachhi, A. Therattil, A. Patil, H. B. Sailor, and H. A. Patil. Analysis of non-linearities in normal *vs.* dysarthric speech for severity-level classification. *submitted in Signal Processing and Communications (SPCOM), IISC, Banglore, India*, 2022.

[43] A. Kachhi, A. Therattil, A. Patil, H. B. Sailor, and H. A. Patil. Dysarthric speech severity-level analysis and classification using teager energy cepstral features. *submitted in INTERSPEECH, South Korea*, 2022.

[44] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering*, vol. 36, pp. 233–247, 2016.

[45] J. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, USA*, volume 1 pp. 381-384, 1990.

[46] J. F. Kaiser. Some useful properties of Teager's energy operators. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis, MN, USA*, volume vol. 3, pp. 149–152, 1993.

[47] M. R. Kamble and H. A. Patil. Analysis of reverberation via Teager energy features for replay spoof speech detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2607–2611, Brighton, UK*, 12-17 May 2019.

[48] M. R. Kamble and H. A. Patil. Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection. In *INTERSPEECH, pp. 646–650*, Hyderabad, India, Sept. 2018.

[49] M. R. Kamble and H. A. Patil. Detection of replay spoof speech using Teager energy feature cues. *Computer Speech & Language*, vol. 65, pp. 101140, January 2021.

[50] R. D. Kent, S. Adams, and G. Turner. Models of speech production. *Contemporary issues in experimental phonetics*, vol. 13, 1976.

[51] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame. Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[52] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *INTERSPEECH, pp.2–6*, Stockholm, Sweden, 2017.

[53] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, et al. Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) pp. 5395–5399, LA, USA*. IEEE, 2017.

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

[55] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King. Attentive filtering networks for audio replay attack detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)pp. 6316–6320*. IEEE, 2019.

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

[57] S. Lefkimmiatis, P. Maragos, and A. Katsamanis. Multisensor multiband cross-energy tracking for feature extraction and recognition. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4741–4744, Las Vegas, NV, USA*, 2008.

[58] 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*, vol. 43,pp. 47–57, 1978.

[59] C. Mackenzie and A. Lowit. Behavioural intervention effects in dysarthria following stroke: communication effectiveness, intelligibility and dysarthria impact. *International Journal of Language & Communication Disorders*, vol. 42, pp. 131–153, 2007.

[60] Madhu R. Kamble and Shekhar Nayak and M. Ali Basha, Shaik and Shakti P. Rath and Vikram Vij and Hemant A. Patil . Teager energy subband filtered features for near and far-field automatic speech recognition. In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 491–496*, Tokyo, Japan, 2021.

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

[62] P. Maragos, J. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing*, vol. 41, pp. 3024–3051(10), 1993.

[63] S. Marcel, M. S. Nixon, and S. Z. Li. *Handbook of Biometric Anti-spoofing*, volume vol. 1. Springer, 2014.

[64] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. Technical report, National Inst of Standards and Technology Gaithersburg MD, 1997.

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

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

[67] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee. Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, pp. 252–265, 2021.

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

[69] P. Maragos, J.F. Kaiser and T.F. Quatieri . On separating amplitude from frequency modulations using energy operators. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–4, San Francisco, California, USA, 1992.

[70] M. S. Paja and T. H. Falk. Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Thirteenth Annual Conference of the International Speech Communication Association,pp. 62-65 OR, USA September 9-13,*, 2012.

[71] A. T. Patil, R. Acharya, A. S. Pulikonda, and H. A. Patil. Energy separation-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection. In *INTERSPEECH, pp. 2898–2902*, Graz, Austria, Sept. 2019.

[72] A. T. Patil, A. Rajul, P. Sai, and H. A. Patil. Energy sepration-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection. In *INTERSPEECH, pp. 2898–2902*, Graz, Austria, Sept. 2019.

[73] H. A. Patil and M. R. Kamble. A survey on replay attack detection for automatic speaker verification (ASV) system. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1047–1053, Honolulu, Hawaii, USA, Nov. 2018.

[74] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. 1^{st} Edition, Pearson Education India, 2015.

[75] D. A. Reynolds. An overview of automatic speaker recognition technology. In *2002 IEEE international conference on acoustics, speech, and signal processing*, volume vol. 4, pp. IV–4072. IEEE, 2002.

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

[77] I. Rodomagoulakis and P. Maragos. Improved frequency modulation features for multichannel distant speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 841–849, 2019.

[78] A. E. Rosenberg. Automatic speaker verification: A review. *Proceedings of the IEEE*, vol. 64, pp. 475–487, 1976.

[79] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease. *The journal of the Acoustical Society of America (JASA)*, vol. 129, pp. 350–367, 2011.

[80] N. J. Shah, M. A. B. Shaik, P. Periyasamy, H. A. Patil, and V. Vij. Exploiting phase-based features for whisper vs. speech classification. In *2021 29th European Signal Processing Conference (EUSIPCO), pp. 21–25*. IEEE, 2021.

[81] S. Sharma, S. Sharma, and A. Athaiya. Activation functions in neural networks. *towards data science*, vol. 6, pp.310–316, 2017.

[82] T. Sorensen, E. Zane, T. Feng, S. Narayanan, and R. Grossman. Cross-modal coordination of face-directed gaze and emotional speech production in school-aged children and adolescents with asd. *Scientific reports*, vol. 9, pp. 1–11, 2019.

[83] Y. Stylianou. Voice transformation: A survey. In *ICASSP, pp. 3585–3588*, Taipei, Taiwan, April 2009.

[84] H. Tak and H. A. Patil. Novel linear frequency residual cepstral features for replay attack detection. In *INTERSPEECH, pp. 726–730*, Hyderabad, India, Sept. 2018.

[85] H. M. Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 599–601, 1980.

[86] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. *Speech Production and Speech Modelling, Springer, pp. 241–261*, 1990.

[87] A. Therattil, P. Gupta, P. K. Chodingala, and H. A. Patil. Teager energy based-detection of one-point and two-point replay attacks: Towards cross-database generalization. *accepted in The Speaker and Language Recognition Workshop (Speaker Odyssey)*, 2022.

[88] A. Therattil, A. Kachhi, and H. A. Patil. Cross-teager energy ceptsral coefficients for dysarthric severity-level classification. *submitted in ISCA Archive*, 2022.

[89] A. Therattil, A. T. Patil, and H. A. Patil. On significance of cross-teager energy cepstral coefficients for replay spoof detection on voice assistants. *submitted in The Speaker and Language Recognition Workshop (Speaker Odyssey)*, 2022.

[90] M. Todisco, H. Delgado, and N. Evans. Constant-Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, vol. 45, pp. 516–535, June 21-24, 2017.

[91] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee. ASVSpoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[92] J. C. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet, and E. Nöth. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of communication disorders*, vol. 76, pp. 21–36, 2018.

[93] R. Vergin and D. O'Shaughnessy. Pre-emphasis and speech recognition. In *Proceedings 1995 Canadian Conference on Electrical and Computer Engineering*, volume 2, pages 1062–1065. IEEE, 1995.

[94] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, vol. 64, pp. 101114, November 2020.

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

[96] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, vol. 66 pp. 130–153, 2015.

[97] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015, pp. 2037–2041, Dresden, Germany.*

[98] Z. Wu, A. Larcher, K.-A. Lee, E. Chng, T. Kinnunen, and H. Li. Vulnerability evaluation of speaker verification under voice conversion spoofing:

the effect of text constraints. In *INTERSPEECH, pp. 950–954*, Lyon, France, August 2013.

[99] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado. Asvspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 588–604, 2017.

[100] K. M. Yorkston, E. A. Strand, and M. R. Kennedy. Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, vol. 5, pp. 55–66, 1996.

[101] 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, pp. 2938–2942,Hyderabad, India*, 2018.

[102] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

[103] W. Zhizheng, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov. ASVSpoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. *INTERSPEECH, Dresden, Germany*, pages 2037–2041, 6-10 September 2015.

[104] J. Zhong, Y. Gan, J. Young, L. Huang, and P. Lin. A new block-based method for copy move forgery detection under image geometric transforms. *Multimedia Tools and Applications*, vol. 76, pp. 14887–14903, 2017.