Development of Countermeasures for Voice Liveness and Spoofed Speech Detection

by

CHODINGALA PIYUSHKUMAR KIRITBHAI
202015002

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

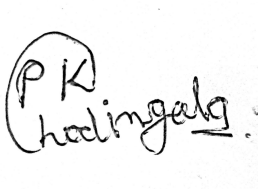
May 2022

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.



Chodingala Piyushkumar Kiritbhai

Certificate

This is to certify that the thesis work entitled "Development of Countermeasures for Voice Liveness and Spoofed Speech Detection" has been carried out by Mr. Chodingala Piyushkumar Kiritbhai for the degree of Master of Technology in Electronics and Communications at *Dhirubhai Ambani Institute of Information and Communication Technology & C.R.Rao Advanced Institute of Applied Mathematics, Statistics and Computer Science* under my supervision.

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

Acknowledgments

"No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude". Hence, I would like to thank all the well wishers for their continuous and excellent support to carry out this work. First, I would like to thank my thesis supervisor Prof. Hemant A. Patil for his guidance, timely suggestions, and motivation to do effective research work. I would like to thank DA-IICT Gandhinagar, for providing resources in the form of books, digital resources, softwares, and a green and peaceful environment that had helped me throughout my thesis work. I would also like to appreciate the Speech Research Lab of DA-IICT, for providing the right platform to nurture young researchers to become professionals in the future.

I would like to extend my sincere thanks to PhD. scholars Mr. Ankur T. Patil, Ms. Priyanka Gupta for their support, guidance, and suggestions throughout my thesis. I would like to thank my friends Ms. Shreya Chaturvedi (Golu), Ms. Aastha Kacchi, Mr. Tejavardhan Reddy, Mr. Lalit, and Mr. Anand Therattil to keep me motivated and help me learn new ideas that assisted me for my overall development. I would also like to acknowledge Mr. Rishabh Pandit for teaching me machine learning techniques. Finally, I would like to extend my profound gratitude to my family members for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. I would like to thank all those whose names have not been called but helped me throughout my studies. Last but not the least, I would like to thank Ms. Motu Patlu to keep me motivated throughout my MTech journey.

Contents

Abstract

An Automatic Speaker Verification (ASV) or voice biometric system performs machine-based authentication of speakers using voice signals. ASV is a voice biometric system which has applications, such as banking transactions using mobile phones. Personal information, and banking details, demand more robust security of ASV systems. Furthermore, the Voice Assistants (VAs) are also known for the convenience of controlling most of the surrounding devices, such as user's personal device, door locks, electric appliances, etc. However, these ASV and VA systems are also vulnerable to various spoofing attacks, such as details, twins, Voice Conversion (VC), Speech Synthesis (SS), and replay. In particular, the user's voice command can be conveniently recorded and played back by the imposter (attacker) with negligible cost. Hence, the most harmful attack (replay attack) of morphing user's voice command can be performed easily. Hence, this thesis aims to develop countermeasure to protect these ASV and VA systems from replay attacks. In addition, this thesis is also an attempt to develop Voice Liveness Detection (VLD) task as countermeasure for replay attack.

In this thesis, the novel Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm (CFCCIF-QESA) feature set is proposed for replay Spoofed Speech Detection (SSD) on ASV systems. Performance of the proposed feature set is evaluated using publicly available datasets such as, ASVSpoof 2017 v2.0 and BTAS 2016. Furthermore, the significance of Delay-and-Sum (DAS) beamformer over state-of-the-art Minimum Variance Distortionless Response (MVDR) for replay SSD on VAs. Finally, the wavelet-based features are proposed for VLD task. The performance of proposed wavelet-based approaches are evaluated using recently released POp noise COrpus (POCO).

Keywords: Automatic Speaker Verification (ASV), Voice Assistants (VAs), Spoofed Speech Detection (SSD), Beamforming, Voice Liveness Detection (VLD).

List of Principal Symbols and Acronyms

AM-FM Amplitude Modulation Frequency Modulation

ASV Automatic Speaker Verification

CFCCIF-QESA Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm

CNN Convolutional Neural Network

CQCC Constant Q Cepstral Coefficients

CWT Continuous Wavelet Transform

DAS Delay and Sum

DCT Discrete Cosine Transform

DET Detection Error Trade-off

EER Equal Error Rate

ESA Energy Separation Algorithm

ESD Energy Spectral Density

GMM Gaussian Mixture Models

IF Instantaneous Frequency

LCNN Light-CNN

LLR Log-Likelihood Ratio

MFCC Mel Frequency Cepstral Coefficients

MI Mutual Information

MVDR Minimum Variance Distortionless Response

POCO POp noise COrpus

ReMASC Realistic Replay Attack Microphone Array Speech Corpus

SS Speech Synthesis

SSD Spoofed Speech Detection

TECC Teager Energy Cepstral Coefficients

TEO Teager Energy Operator

VC Voice Conversion

VLD Voice Liveness Detection

List of Tables

List of Figures

CHAPTER 1
Introduction

1.1 Motivation

Identity recognition incorporates various biometric traits, such as voice, fingerprint, iris, face, palmprint. Among these, voice as a biometric trait is emerging due to its naturalness and ease of production. To that effect, it has led to the development of speaker identification and verification systems. In particular, Automatic Speaker Verification (ASV) systems are also called as voice biometric systems. However, recent parallel developments in several speech technology applications, such as voice conversion, synthetic speech, and high quality microphones and speakers have paved the way to breach (attack) ASV systems by presenting the fake voice samples of the claimed identity, which are known as *spoofing attacks*. These spoofing attacks are categorized, such as impersonation by twins [10], Speech Synthesis (SS) [11], Voice Conversion (VC) [12], and replay [13]. Among these known spoofing attacks, replay attacks are the easiest to mount, however, difficult to detect due to the availability of high quality recording and playback devices [14]. These attacks make the ASV system vulnerable and questions the applicability of the ASV system in financial and privacy applications, such as banking and voice assistants. These spoofing attacks can be overcome by either developing the robust ASV system or implementing the separate countermeasure system, which assist the ASV to detect the spoofing attacks. However, earlier approach will diminish the performance of the ASV system as there is a trade-off between the performance of the ASV system and its robustness against the spoofing attacks. This trade-off exists due to the fact that development of the robust architecture (either feature set or classifiers) affects the speaker-specific characteristics. Hence, ASV research community focused upon the developing efficient countermeasure (CM) systems against spoofing attacks. To that effect, the development in the ASVSpoof challenge campaigns emerged through the discussion in first edition of special session in *Spoofing and Countermeasures for ASV*, held during

INTERSPEECH 2013 [15]. These discussions helped to provide the common platform and procedures for implementation of the CMs with statistically meaningful datasets, protocols, and evaluation metrics and emerged as ASVSpoof-2015, -2017, -2019, and -2021 challenges [1, 16–18]. The ASVSpoof-2015 challenge focused on developing the CM systems for the spoofed speech signals generated from well established text-to-speech (TTS) and voice conversion (VC) techniques. However, replay speech signals are easy to generate with the help of easily available high quality microphones and speakers. To address the vulnerability of the ASV system against replay spoofing attacks, ASVSpoof-2017 challenge was designed to develop CMs against real replay speech signals. Furthermore, the ASVSpoof-2019 challenge introduced two scenarios, namely, logical access (LA) and physical access (PA), where LA addresses the spoofing attacks generated by the TTS and SS, and PA addresses the replay spoofing attacks. The LA scenario in ASVSpoof-2019 challenge edition considers the TTS- and VC-based speech signals generated by neural network-based vocoders. Whereas, the replay speech signals are simulated using a range of real replay devices and carefully controlled acoustic conditions.

Furthermore, along with the ASV systems, the Voice Assistants (VAs) are also known for the convenience of controlling most of the surrounding devices, such as users' personal devices, door locks, electric appliances, etc. [19]. There are several VA systems available, such as Apple Siri, Google Assistant, Microsoft Cortana, Samsung Bixby, etc. However, these VAs are also highly vulnerable against spoofing attacks, similar to ASV systems. Although ASV and VAs seem similar, there is a significant difference, such as ASVs are designed for mono-channel audio and near-field speech, while VAs are designed multi-channel audio and far-field speech. To that effect, Realistic Replay Attack Microphone Array Speech Corpus (ReMASC) is designed to develop CMs for VAs [20].

Furthermore, to improve the security of ASV and VA systems against these spoofing attacks, the Voice Liveness Detection (VLD) system is used [21, 22]. In this context, a 'liveness' detection corpus called as the POp noise COrpus (POCO) has been released in 2020 to allow research on development of robust VLD systems [5, 21]. One of the cues of liveness in a speech signal is the presence of *pop noise* in a live (genuine) speech signal. Pop noise is a short-time distortion in a speech signal, which is caused by a burst of air on the microphone originating from a live speaker's mouth [23]. Signals that are known to spoof ASV systems, such as synthetic speech and replayed speech, fail to reproduce the pop noise as strongly as a live speech signal [5, 24], of course, with the assumption that spoofed speech is not recorded with *wiretapping*. Pop noise is found in live speech

as sudden bumps of strong energy within duration ranging between 20 ms and 100 ms [21].

1.2 ASV with SSD System Architecture

An ASV system consists of two parts, namely, speaker verification (SV) and spoof detection. SV deals with verifying a person's identity to a known identity while spoof detection, on the other hand, identifies the naturalness of the input speech signal. A general spoof detection strategy consists of extracting discriminative features or representations of the input speech, which, can be used by the trained model to classify it as genuine or fake utterance. The Figure 1.1 shows the architecture of speaker verification system along with SSD system. Here, the SSD system identify the naturalness of the input speech and reject if it is spoofed one, whereas if the claimed speech is natural then it will feed it to ASV or VA system for further verification.

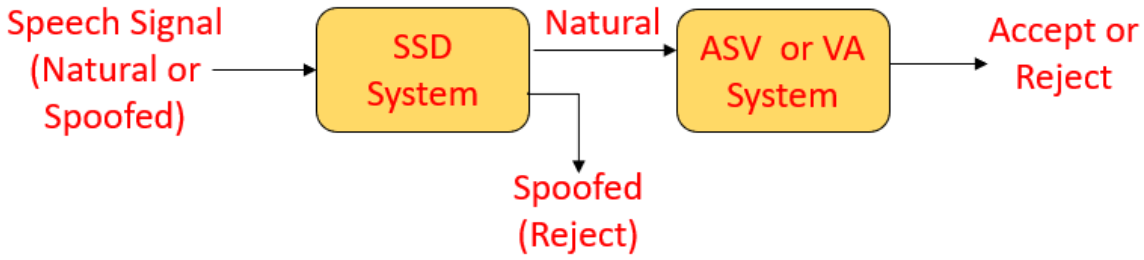


Figure 1.1: Block diagram of basic SSD system with ASV and VA system.

1.3 Application of Spoofed Speech Detection (SSD) and Voice Liveness Detection (VLD) System

Some applications of the SSD and VLD systems are as follows:

- As the availability of high quality recording and playback devices are increasing, the user's voice command can be conveniently recorded and played back by the imposter (attacker) to get unauthorized access of ASV and VA systems. Hence, the most harmful attack of morphing user's voice command can be performed easily. Hence, to protect our ASV and VA systems from these types of spoofing attacks, we can employ VLD system, to confirm the presence of live speaker.

- The use of voice biometric system has been increasing intensively in the field of banking and financing sector. Hence, these biometric systems should be robust enough to prevent the spoofing attacks. To that effect, the development of VLD and SSD systems are utmost needed.
- In order to verify the originality of a speech signal or a recording, SSD systems are used in forensic departments to verify the legitimacy of a speech signal.

1.4 Contributions of The Thesis

1.4.1 CFCCIF-QESA Feature Set

In the past, auditory transform-based as well as Instantaneous Frequency (IF)-based features have been proposed for replay SSD. In this context, IF has been estimated either by derivative of analytic phase via Hilbert transform, or by using high temporal resolution Teager Energy Operator (TEO)-based Energy Separation Algorithm (ESA). However, excellent temporal resolution of ESA comes with lacking in using relative phase information, and vice-versa. To that effect, we propose novel Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm (CFCCIF-QESA) features, with excellent temporal resolution as well as relative phase information. CFCCIF-QESA is designed by exploiting relative phase shift to estimate IF, without estimating phase explicitly from the signal.

1.4.2 Significance of DAS *vs.* MVDR Beamformer for Replay SSD on VAs

Voice Assistants (VAs) are becoming more useful in daily life and hence, the safety of VAs from various spoofing attacks is crucial. To that effect, we analyze the significance of *delay and sum* (DAS) beamforming technique over the state-of-the-art Minimum Variance Distortionless Response (MVDR) beamformer for replay SSD for VAs. In particular, DAS is known to suppress the additive noise component and retains the characteristics of replay mechanism and hence, DAS can be exploited for replay SSD in VAs. On the contrary, MVDR beamforming is proved to be the efficient beamformer for far-field speech recognition application, however, it suppresses the additive noise along with the characteristics of replay mechanism. Hence, MVDR is not suitable choice for replay SSD in VAs.

1.4.3 Wavelet-Based Features

Given the attacker's freedom of using any spoofing attack, there is a need to explore liveness detection approaches that can classify a live speech from all the various spoofed speeches. To that effect, in this thesis work, the Morlet wavelet-based approach for Voice Liveness Detection (VLD) is proposed. We use acoustic cues of pop noise to discriminate a live speech signal from a spoof speech. Pop noise is present in live speech signals at low frequencies, caused by human breath reaching at the closely-placed microphone.

1.5 Organization of the Thesis Work

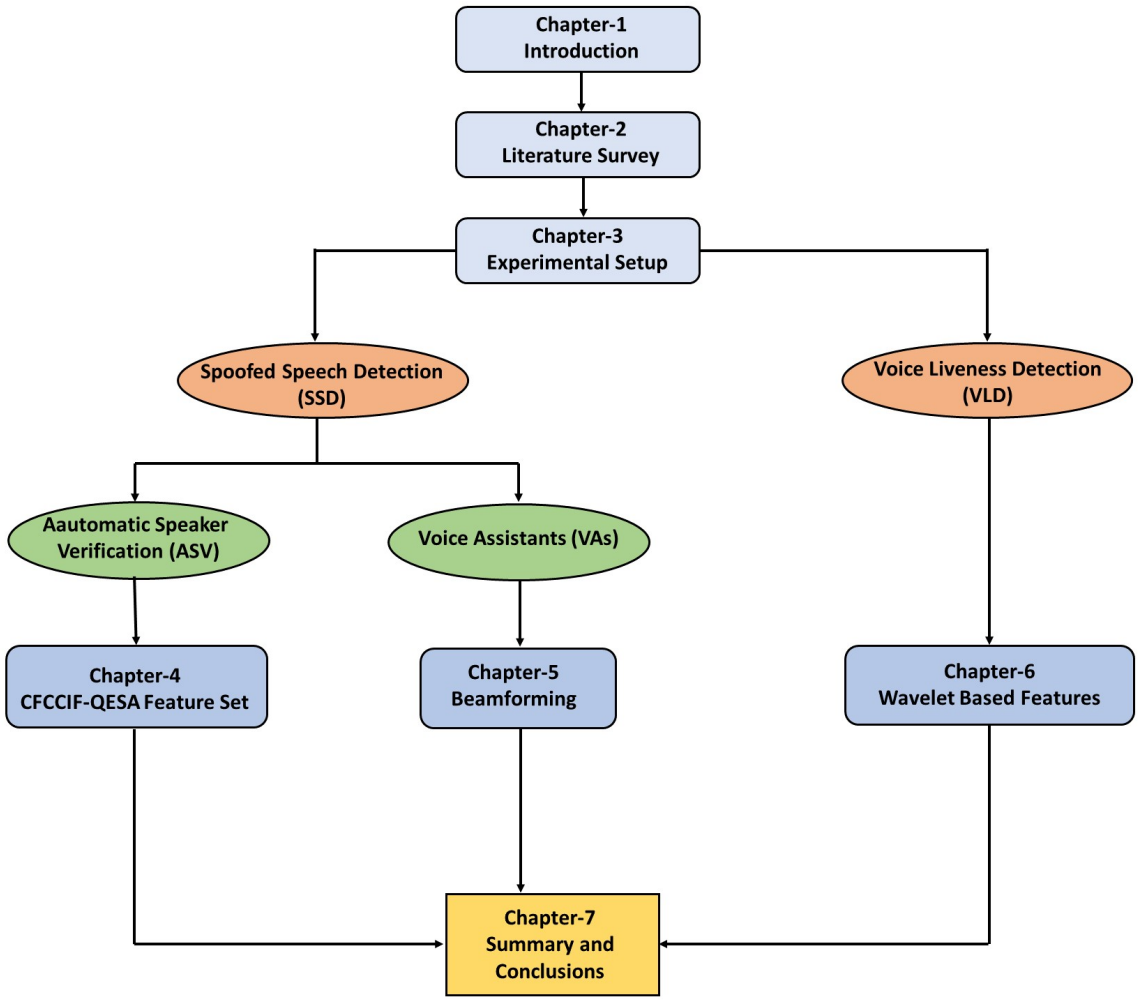


Figure 1.2: Flowchart of the Thesis.

Chapter-2 presents the literature survey of the previous research done in the field of SSD and VLD systems for ASV. This chapter illustrates the systems devel-

oped for VLD and SSD.

Chapter-3 discusses the details of basic SSD and VLD systems along with various processing required for executing these tasks. Furthermore, this chapter illustrates the performance measures used for evaluation of systems. The details of datasets and classifier are also described in this chapter.

Chapter-4 presents the proposed Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm (CFCCIF-QESA) features for replay SSD task. Using this feature set, we are supposed to estimate the Instantaneous Frequency (IF) using QESA, which incorporates the magnitude as well as phase information of the speech signal.

Chapter-5 discusses the significance of the DAS beamformer over MVDR for replay SSD on VAs. The speech signal of replay consists of the replay mechanism characteristics and spatial noise. The replay mechanism characteristics are convolved with the original (i.e., genuine) signal, whereas spatial noise is additive in nature. For replay SSD, the approach should be developed such that it should suppress the spatial noise and retain the characteristics of replay mechanism (i.e., reverberation). This can be achieved by selecting the DAS beamformer as opposed to the MVDR.

Chapter-6 describes the use of Morlet wavelet-based features for pop noise detection. With respect to Heisenberg's uncertainty principle in signal processing framework, wavelet-based approach offers improved resolution in time and frequency as compared to the STFT-based method.

Chapter-7 gives the overall summary of this thesis work along with the application and the limitations of this work. In addition, this chapter presents the potential future research directions.

1.6 Chapter Summary

This chapter presented spoof and voice liveness detection problems on ASV systems, and also described the key motivation for this thesis work. Furthermore, the applications of SSD and VLD systems are explained. Finally, the overall organization of the thesis work is explained. The next chapter describes the literature survey done on SSD and VLD systems, along with its limitations.

CHAPTER 2
Literature Survey

2.1 Introduction

This chapter describes the previous efforts made in the direction of making robust countermeasures for SSD. In the beginning, the chapter presents literature review of previous work done in the direction of replay SSD and VLD for ASV and VA systems. In particular/, few notable previous studies are discussed, which are done on ASVSpoof 2017 v2.0, BTAS 2016, and ReMASC datasets for replay SSD tasks. In the latter part, the chapter describes few studies reported for pop noise detection. This literature search presented in this paper helps to position this thesis work in the history of this research problem.

2.2 Literature Review

2.2.1 Replay SSD

ASV systems have been used for various applications, such as banking transactions and access to systems associated with classified information. It is used to grant access to only an enrolled set of speakers (users). All the remaining speakers are treated as non-genuine or imposter speakers. Nevertheless, some impostors deliberately attempt to get unauthorized access to the ASV system. These deliberate attempts made by the impostor (i.e., attacker) are called as spoofing attacks on ASV. Furthermore, an ASV system becomes robust by suppressing the effect of recording and transmission channel information, acoustic noise, etc. This robustness leads to vulnerability of an ASV system to spoofing attacks. Impersonation, twins, voice conversion (VC), speech synthesis (SS), and replay are the possible spoofing attacks on ASV systems. Out of these spoofing attacks, replay attacks are the easiest to mount but difficult to detect due to the availability of high quality recording and playback devices [14]. In order to develop robust

countermeasures to detect spoofed speech, the first special session on *Spoofing and Countermeasures for ASV* was held during INTERSPEECH 2013 [14, 25]. Details of various vulnerabilities on ASV systems and their respective countermeasures (CMs) were presented in [14]. The need for standard datasets, protocols, and performance evaluation metrics in this special session led to the ASVSpoof 2015 Challenge organized during INTERSPEECH 2015. This challenge concentrated on developing several CMs against SS and VC spoofs using various kinds of feature extraction algorithms on a standard statistically meaningful ASVSpoof 2015 dataset [16, 26–30]. The CMs in this challenge were based on signal processing-based techniques to develop feature sets, and Gaussian Mixture Model (GMM) as pattern classifier for the two-class problem of genuine *vs.* spoof speech detection (SSD). Among the various submissions by the participants, some notable submissions were based on various feature sets, such as Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) (which was the *winner* system during ASVSpoof 2015 challenge [31]), Linear Frequency Cepstral Coefficients (LFCC) [32], and Constant-Q Cepstral Coefficients (CQCC) [32, 33]. Furthermore, in the ASVSpoof 2017 challenge, the focus was exclusively on replay SSD [34–36], whereas in the ASVSpoof 2019 challenge, the focus was on synthetic or simulated replay (also called Physical Access (PA)) SS and VC-based attacks (called as Logical Access (LA)). Lastly, the most recent challenge is the ASVSpoof 2021 challenge with three tracks, namely, LA, PA, and DeepFake detection [37].

After successful completion of ASVSpoof 2015 challenge, which addresses SS and VC type spoofing attacks, the BTAS 2016 challenge was organized which addresses SS, VC, and replay type attacks. Here, a few notable contributions for detection of these three spoofing attacks using BTAS 2016 dataset are shown in Table 2.1. The study in [38], reports the significance of Long Short-Term Memory (LSTM) as classifier with deep neural network-based and CQCC features over the other neural network based systems and CQCC-GMM baseline system for SSD. This study shows that, combination of deep features with CQCC and LSTM as back-end classifier system performs well for replay SSD. In [39], the Teager Energy Operator (TEO)-based features are proposed to compute the running estimate of energy cues for SSD task. In particular, the Teager Energy Cepstral Coefficients (TECC) are extracted using linearly-spaced Gabor filterbank to extract narrowband filtered signals. However, the obtained running energy via TECC is estimated energy and hence, the Enhanced Teager Energy Cepstral Coefficients (ETECC) is proposed in [40] to calculate exact running energy in signal for SSD.

Here, few notable contributions for replay SSD using ASVSpoof 2017 v2.0

database are shown in the Table 2.1. During this challenge, the study reported in [41] shows the relevance of neural network-based classifiers using spectrographic features for replay SSD. In particular, the authors used Convolutional Neural Network (CNN), Light-CNN (LCNN), and ResNet-based classifiers along with spectrogram as features for replay SSD. Next, in [41], the authors employ ResNet as classifier along with Constant Q Cepstral Coefficient (CQCC) features for replay SSD. Furthermore, in this study, the performance of SSD system was improved via score-level fusion of proposed system with the CQCC-GMM system. In the later year, the similarity between genuine and spoofed speech in replay SSD task is detected via deep Siamance features along with GMM is proposed in [42]. In particular, this proposed system performed well as compared to end-to-end deep neural models for replay SSD. The study in [38], reports the significance of Long Short-Term Memory (LSTM) as classifier with deep neural and CQCC features over the other neural network-based systems and CQCC-GMM baseline system for replay SSD. This study shows that, combination of deep features with CQCC and LSTM as back-end classifier system performs well for replay SSD. In [34], a comparative study for replay SSD is reported between various existing feature sets, such as Linear Frequency Cepstral Coefficients (LFCCs), CQCC, Mel Frequency Cepstral Coefficients (MFCCs), Inverted Mel Frequency Cepstral Coefficients (IMFCCs), Linear Prediction Cepstral Coefficients (LPCCs), Rectangular Filter Cepstral Coefficients (RFCCs), Subband Spectral Centroid Frequency Coefficients (SCFCs), Subband Spectral Flux Coefficients (SSFCs), and Subband Spectral Centroid Magnitude Coefficients (SCMCs). In addition, the authors performs cross-database experiments between BTAS 2016 and ASVSpoof 2017 using above listed feature sets to find the consistent performing replay SSD system. In [43], the authors proposed High Frequency Cepstral Coefficients (HFCC) features along with Deep Neural Network (DNN) and Support Vector Machine (SVM) as back-end classifier. Furthermore, the performance of proposed system is improved via score-level fusion with CQCC. In later year, the study reported in [44], achieves 0% EER on both development and evaluation sets of ASVSpoof 2017 v2.0 dataset using modified group delay function (MGDCC)-based features along with ResNet classifier.

For SSD task, in [31], an Auditory Transform (AT)-based Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency (CFCCIF) feature set was proposed. It was based on cochlear filter and IF-based information. To that effect, IF is estimated conventionally from the analytic phase denoted via the Hilbert transform (HT) of the underlying real signal [45]. However, estimating IF from

this approach is computationally expensive. Moreover, the resolution of HT in time-domain is poor, as it requires a block (frame) of speech data [46]. To address this issue, in [47] the authors proposed Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Energy Separation Algorithm (CFCCIF-ESA) feature set which uses Teager Energy Operator (TEO)-based Energy Separation Algorithm (ESA) [48] to estimate IF with high time resolution for replay SSD task [49]. Due to the use of TEO in estimation of IF, CFCCIF-ESA utilizes only the amplitude information of the signal for replay SSD. Moreover, due to absence of HT, it does not contain the quadrature-phase component of the signal. Therefore, in order to incorporate both the advantages, i.e., excellent time resolution of TEO and having quadrature-phase component via HT, in this thesis we proposed Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature based Energy Separation Algorithm (CFCCIF-QESA)feature set.

Furthermore, as the use of VAs increases, the vulnerability of these personal devices are increasing. Hence, the imposters are mounting similar spoofing attacks as ASV, on VAs to get unauthorized access of these devices. Although ASV and VAs seems similar, there is a significant difference, such as ASVs are designed for mono-channel audio and near-field speech, while VAs are designed for multichannel audio and far-field speech. To that effect, recently Realistic Replay Attack Microphone Array Speech Corpus (ReMASC) is designed to develop CMs for VAs [20]. To that effect, in [50], the authors proposed Cross-Teager Energy Operator (CTEO), which select the optimal channel from a multichannel input based on maximum cross-energy computation. Here, via maximizing the cross-energies, we can identify the distortions added due to intermediate devices in replay speech signal. Hence, the proposed system performs well as compared to baseline CQCC system.

2.2.2 VLD System

As we have discussed above, ASV systems are prone to spoofing attacks, such as VC, SS, and replay. Out of these, the replay attack involves the least amount of technological effort and hence, it is the easiest attack to mount on an ASV system. The ASV challenges aimed to develop robust countermeasures against replay attacks under various configurations of recording and the playback devices. However, the perspective of Voice Liveness Detection (VLD) has begun only recently, when the POCO dataset for VLD was released in late 2020 [5, 21]. The POCO dataset relies on *pop noise* cues for detection of live (i.e., genuine) speech. Pop noise is generated due to the human breath reaching the microphone. Pop noise

Table 2.1: Results (in % EER) from the literature of ASVSpoof 2017, BTAS 2016, and ReMASC datasets on replay SSD for ASV and VA systems.

Dataset	Authors	Feature Sets	Classifier	% EER	
				Dev	Eval
ASVSpoof 2017 v2.0 (For ASV System)	Weicheng Cai *et al.* [41]	CQCC	ResNet	10.25	22.39
	Kaavya Sriskandaraja *et al.* [42]	Siamese Embedding Features	GMM	-	6.40
	Galina Lavrentyeva *et al.* [51]	Spectrogram	LCNN+CNN+RNN	3.95	6.73
			LCNN	4.53	7.37
	Lian Huang *et al.* [38]	CQCC+deep features	LSTM	3.13	8.28
	Roberto Font *et al.* [34]	SCMCs	GMM	9.32	11.49
	Parav Nagarsheth *et al.* [43]	HFCC+CQCC	DNN-SVM	7.6	11.5
	Lian Huang *et al.* [52]	CQCC	LSTM	3.62	9.56
	Francis Tom *et al.* [44]	MGDCC	ResNet	0	0
BTAS (2016) (For ASV System)	Lian Huang *et al.* [52]	CQCC+deep features	LSTM	0.09	0.93
	Madhu Kamble *et al.* [39]	TECC	GMM	2.25	4.51
	Ankur Patil *et al.* [40]	ETECC	GMM	1.50	2.95
ReMASC (For VAs)	Rajul Acharya *et al.* [50]	CTECC	GMM	16.46	15.93
		CQCC	GMM	20.57	23.31

is a distortion of short duration, which is caused by a burst of air on the microphone originating from a live speaker's mouth [23]. Due to the proximity of the genuine/live speaker to the microphone, the pop noise is captured at the microphone. However, in the case of a replay attack, the attacker records the speech signal from a considerably large distance in order to be discreet. In this case, the pop noise is absent or weakly captured in the recorded speech signal and thus, making pop noise as characteristics of live speech.

Before POCO dataset was released, only a few studies had been reported for VLD, such as low frequency-based single channel detection and subtraction-based pop noise detection in [21], pop noise-based VLD for smartphones in [53], and phoneme-based pop noise detection in [54]. However, since the release of the POCO dataset, some work has been done in the direction of VLD as shown in Table 2.2. The baseline system, which introduced the POCO dataset, uses Short-Time Fourier Transform (STFT)-based features for detection of pop noise [5]. It has been found that pop noise occurs in low frequency regions of typically ≤ 40 Hz [5]. Pop noise is a short-time distortion in a speech signal, which is caused by a burst of air on the microphone originating from a live speaker's mouth [23]. Signals which are known to spoof ASV systems, such as synthetic speech and replayed speech, fail to reproduce the pop noise as strongly as a live speech signal [5, 24], of course, with the assumption that spoofed speech is not recorded with

wiretapping. Pop noise is found in live speech waveforms as sudden bumps of strong energy within duration ranging between 20 ms and 100 ms [21]. To detect pop noise, various techniques have been proposed in the literature, such as STFT [55], Constant Q-Transform (CQT) [56], and bump wavelet-based features [57]. In particular, the approach of using Continuous Wavelet Transform (CWT) was first proposed in [57], wherein a bump wavelet was used.

Table 2.2: Results (in % Accuracy) from the literature of POCO dataset on pop noise detection for VLD.

Dataset	**Author**	**Feature Sets**	**Classifier**	**Frequency Range**	**% Accuracy**
POCO	Shrishti Singh *et al.* [58]	Spectral Root Smoothing	GMM	0-11025 Hz	69.79
	Sidhhant Gupta *et al.* [55]	STFT	CNN	0-40 Hz	80.51
	Priyanka Gupta *et al.* [57]	Bump Wavelet Based CWT	CNN	0-40 Hz	80.19

2.3 Chapter Summary

This chapter describes literature review of SSD and VLD systems in brief. Firstly, few notable studies has been discussed for replay SSD using ASVSpoof 2017 v2.0 and BTAS 2016 datasets following by few top studies reported for replay SSD on VAs using ReMASC dataset. Later, the brief literature survey for liveness detection using POCO dataset are discussed. In the next chapter, we will discuss the experimental set used in this thesis work.

CHAPTER 3
Experimental Setup

3.1 Introduction

In this chapter, we present an overview of corpora used in this work for developing the CMs against the spoofing attacks. In particular, the details of ASVSpoof 2017 v2.0, BTAS 2016, ReMASC, and POCO datasets along with their partitions, recording conditions, etc. are given. Furthermore, the details of architecture of various classifiers used in this study are given. In addition, the details of evaluation metrics and data fusion techniques used in this study.

3.2 Datasets Used

3.2.1 ASVSpoof-2017 Challenge Dataset

This dataset was released for ASVSpoof 2017 challenge organized during INTERSPEECH-2017, and later it was publicly available [59]. However, data anomalies, such as period of silence and zero value samples have been noticed by the challenge organizers and then these anomalies are addressed in the second version of the dataset, which is known as ASVSpoof-2017 Version 2.0 dataset [1][1]. In this dataset, genuine utterances are selected from the RedDots corpus, which is designed for text-dependent ASV using *ten* prompt sentences [60]. Replay spoof signals are generated in 177 sessions using various acoustic environments and heterogeneous devices. The standard partition of the dataset into training, development, and evaluation is done as shown in Table 3.1 [1]. There are 61 distinct replay configurations which are a combination of a playback device, a recording device, and an acoustic environment.

[1]The ASVSpoof 2017 V2.0 dataset along with protocols and extended metadata is available online at *https://datashare.ed.ac.uk/handle/10283/3055 {Last access: 18, May 2022}*

Table 3.1: Statistics of the ASVSpoof 2017 dataset for the environment-independent case. After [1].

Subset	# Spk	Utterances		Environments
		Genuine	Spoof	
Train	10	1507	1507	E3, E6
Dev	8	760	950	E3, E5, E6
Eval	24	1298	12008	E1 - E7
Total	42	3565	14465	

E1: Anechoic Room, E2: Analog Wire, E3: Balcony, E4: Canteen, E5: Home, E6: Office, E7: Studio, Spk: Speaker

Table 3.2: Distribution of spoof speech utterances among the environments in ASVSpoof 2017 dataset. After [1].

Environment	**# Utterances**	**Environment**	**# Utterances**
Anachoic	748	Canteen	3517
Analog Wire	543	Office	7565
Balcony	1184	Studio	342
Home	570	-	-

3.2.2 Biometrics: Theory, Applications, and Systems-2016 (BTAS-2016)

This dataset was released during the speaker anti-spoofing competition during IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS-2016) [3]. It considers all the major types of spoofing attacks, namely, replay, SS, and VC. [2]BTAS-2016 dataset uses AVSpoof dataset [2]. Genuine utterances in BTAS-2016 dataset are recorded from 44 subjects, which consists of 31 males and 13 females. The recording is performed in 4 sessions over the period of 2 months, with varying recording setups and environmental conditions. Three types of recording devices, namely, laptop using microphone AT2020USB+, Samsung Galaxy S4 phone, and iPhone 3GS are utilized for the genuine speech signal recording with 3 types of sample recordings: (1) reading part of 10 or 40 predefined sentences read by subjects (read), (2) pass-phrases part of 5 short prompts read by subjects (pass-phrases), and (3) free speech part of a free speech about any topic for 3 to 10 minutes (free). The details of the BTAS-2016 dataset recordings w.r.t. the session and recording type is shown in Table 3.3. The statistics of the dataset w.r.t. replay configuration is shown in Table 3.4. It can be observed that the training and development set consists of the similar kind of spoofing attack

[2]The BTAS 2016 dataset is available at *https://www.idiap.ch/en/dataset/avspoof {Last access:18, May 2022}*

algorithms. Hence, known attacks are present in the development set. However, the test set consists of two unseen replay attacks, which are known as unknown attacks.

Table 3.3: Statistics of the BTAS-2016 dataset w.r.t. the session and recording type. After [2].

	Session 1	Session 2-4	Total
read	10 sentences	40 sentences	25.96 hours
pass-phrases	5	10	4.73 hours
free	$\geq$ 5 min	$\geq$ 3 min	38.51 hours

Table 3.4: Number of utterances in BTAS-2016 dataset. Acronyms in this Table stands for the following terms: SS- Speech Synthesis, VC- Voice Conversion, RE- Replay, LP- Laptop, PH1- Samsung Galaxy S4 phone, PH2- iPhone 3GS, PH3- iPhone 6S, HQ- High Quality Speakers. After [3].

	Train	Dev	Test
genuine	4973	4995	5576
spoof	38580	38580	44920
SS-LP-LP	490	490	560
SS-LP-HQ-LP	490	490	560
VC-LP-LP	17400	17400	19500
VC-LP-HQ-LP	17400	17400	19500
RE-LP-LP	700	800	800
RE-LP-HQ-LP	700	800	800
RE-PH1-LP	700	800	800
RE-PH2-LP	700	800	800
RE-PH2-PH3	-	-	800
RE-LPPH2-PH3	-	-	800

3.2.3 ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems

ReMASC corpus is specifically designed to develop the CMs for VAs [4].[3]There are important differences between ASV and VAs, primarily, the distance between the speaker and microphone is larger in VAs. Furthermore, VAs utilize a microphone array as opposed to the single microphone in ASV. In the ReMASC dataset, 132 voice commands are used. These voice commands consist of 273 unique words for phonetic diversity. The number of speakers in the dataset are 50, out of which 22 are female speakers and 28 are male speakers. Furthermore, out of 50, 36 speakers are native speakers of English language, 12 are Chinese native speakers, and 2 are Indian speakers. The speech data is collected for 4 systems, details of

[3]This dataset is publicly available at *https://github.com/YuanGongND/ReMASC {Last access:18, May 2022}*

which are shown in Table 3.5. Furthermore, to study the effect of recording device in replay attack, one low quality (iPod Touch (Gen5)) and one high quality recorder (Tascam DR-05) is used. However, it is observed that even with Tascam DR-05, channel and background noise are unavoidable. To that effect, for additional replay source recordings, Google Text-To-Speech (TTS) is used, which is free from channel and background noise. For playback, 4 devices are used: A) Sony SRSX5, B) Sony SRSX11, C) Audio Technica ATH-AD700X headphone, and D) iPod Touch. Moreover, an additional playback device is used in the vehicular environment as the built-in vehicular audio system. The ReMASC data is recorded in 4 types of environments, namely, outdoor environment, vehicle environment, indoor environment-1, and indoor environment-2. The statistics of the dataset along with corresponding environments is shown in Table 3.6.

For this dataset, standard partition, protocols, and evaluation metrics, are not provided by the dataset organizers. However, few architectures developed using various dataset configurations of ReMASC dataset can be studied in [40, 61, 62]. In [40], the authors proposed Enhanced Teager Energy Cepstral Coefficients (ETECC) features for replay SSD on VAs. The study in [61] shows the significance of Spectral Root Cepstral Coefficients (SRCC) features for replay SSD using GMM as classifier. Furthermore, the novel neural network-based model is proposed in [62] to improve the replay SSD for VAs.

Table 3.5: Microphone array settings for ReMASC dataset. After [4].

Device	**Sampling Rate (in Hz)**	**Bit depth**	**Number of channels**
Amlogic 113X1	16000	16	7
Respeaker 4 Linear	44100	16	4
Respeaker V2	44100	32	6
Google AIY	44100	16	2

Table 3.6: Statistics of the ReMASC dataset w.r.t. various acoustic environments. After [4].

Environment	# Subjects	# Genuine	# Spoof
Outdoor	12	960	6900
Vehicle	10	3920	7644
Indoor-1	23	2760	23104
Indoor-2	10	1600	7824

3.2.4 POp noise COrpus (POCO)

A significant amount of work has been done in the SSD literature. However, detection of live speech has only been paid attention to recently, by using the recent standard corpora, POp noise COrpus (POCO) [5] [4].For liveness detection of speech, pop noise is utilized as characteristics of live speech. Pop noise is produced due to the breathing effects captured by the microphone. If microphone in ASV system is *assumed* to be placed close to the genuine/live speaker, then it is able to capture the pop noise effectively. Therefore, pop noise becomes a suitable acoustic feature for distinguishing a live speech from a spoof (especially replayed) speech signal. To that effect, the POCO dataset is developed to investigate the liveness feature of ASV.

The POCO dataset consists of speech recordings of 66 speakers (32 male and 34 female), aged from 18 to 61 years, with varying levels of English language fluency and accent. The dataset is recorded with 22050 Hz sampling frequency and a bit-depth of 16-bits. The dataset is organized into three parts, namely, RC-A, RP-A, and RC-B. These parts differ from each other in number of microphones, type of microphone(s) used, and presence/absence of pop filter. The details of these 3 parts are given in Table 3.7. The subset RC-A represents live speaker recordings having pop noise. The subset RP-A consists of emulated scenario of spoofed speech by using pop filter to eliminate/diminish pop noise. While RC-A and RP-A consists of speech data captured by a single microphone, the subset RC-B consists of speech data captured by an array of 15 microphones. Like the RC-A subset, the RC-B subset also doesn't use pop filter and, hence, corresponds to live speech. Speech signals in RC-B set are recorded in 3 settings w.r.t. speaker-microphone distances, namely, 5 cm, 10 cm, and 20 cm. The effect of human breath on the microphone depends on the uttered phoneme type. Thus, the POCO dataset is collected such that it consists of speech recordings of 44 words corresponding to 44 phonemes in the English language, as shown in Table 3.8.

For each of the recording setting (RC-A, RP-A, RC-B (5 cm), RC-B (10 cm), and RC-B (20 cm)), each word shown in Table 3.8 was repeated 3 times by every speaker. Furthermore, in the case of RC-B setting, where multiple microphones were used, all the microphones were tuned independently so that the maximum volume remained below the threshold of -6 dB. The dataset is not partitioned into training, development, and evaluation subsets. Thus, experiments using this dataset can be conducted by considering non-overlapping training, and testing subsets.

[4]The POCO dataset can be found at *https://github.com/aurtg/poco {Last access:18, May 2022}*

Few architectures on voice liveness detection using POCO dataset, can be studied in [58, 63, 64]. In [63], due to the low frequency constraint, the authors shows the low frequency wavelet transform-based features using CNN as classifier. The study in [58] shows the significance of modified group delay features for pop-noise detection with GMM-based classifier. In [64], the authors shows significance of CQT-based features for voice liveness detection using SVM-based classifier.

Table 3.7: The three subsets of POCO dataset. After [5].

Subset	Microphone Name	Microphone Frequency Response (in Hz)	Microphone Directionality	Number of Microphones	Distance of speaker from the microphone (in cm)	Pop Filter
RC-A	audio-technica AT4040	20 - 20,000	Cardoid	1	10	No pop filter used
RP-A	audio-technica AT4040	20 - 20,000	Cardoid	1	10	TASCAM TM-AG1
RC-B	audio-technica AT9903	30 - 18,000	Omnidirectional	15	5, 10 and 20	No pop filter used

Table 3.8: The 44 words spoken 3 times each by each speaker. After [5].

44 words in the POCO dataset						
about	arm	laugh	bird	bug	busy	chair
chip	dad	division	end	exaggerate	fat	five
funny	gun	his	honest	hop	join	kit
leather	live	monkey	open	paw	pay	pin
pink	quick	summer	sham	shout	sit	spider
steer	run	thong	tip	tourist	who	wolf
you	be					

3.3 Classifiers Used for SSD

In this work, binary classification is done using three types of classifiers, namely, Gaussian Mixture Model (GMM), Convolutional Neural Network (CNN), and Light-CNN (LCNN). While our primary emphasis is on the improved performance due to the proposed quadrature-based feature set, we also trade it with the effect of different statistical and deep learning based classifiers. The details of each of the classifiers is explained in this sub-Section.

3.3.1 Gaussian Mixture Models (GMM)

GMM is a parametric model, which is represented as a weighted sum of Gaussian component probability densities. In particular, the parameters of a GMM are estimated using the Expectation Maximization (EM) algorithm iteratively using the

training data. The GMM can be represented as a weighted sum of N component Gaussian densities, which is given by [65]:

$$p(v/\lambda) = \sum_{i=1}^{N} w_i g(v/\mu_i, \Sigma_i), \tag{3.1}$$

where v represents a D-dimensional continuous feature vector. Further, the mixture weights are represented by w_i, i=1, 2, ..., N, and g(v/μ_i,Σ), i=1,2, ..., N are densities of the Gaussian mixture components. We wish to estimate parameters of GMM, when the training vector and GMM configuration are given. This estimation is done using EM algorithm which starts with λ (initial model), to estimate $\bar{\lambda}$ (new model), such that P(V/$\bar{\lambda}$) $\geq$ P(V/λ). This estimated new model then becomes the initial model for estimation of the next model. This process is repeated iteratively till it reaches a convergence threshold. The expression for estimate various parameters of GMM using EM algorithm is [65]:

(1) **Mixture weights:**

$$\bar{w_i} = \frac{1}{T} \sum_{t=1}^{T} P_r(i/x_t, \lambda), \tag{3.2}$$

(2) **Means:**

$$\bar{\mu_i} = \frac{\sum_{t=1}^{T} P_r(i/x_t, \lambda) x_t}{\sum_{t=1}^{T} P_r(i/x_t, \lambda)}, \tag{3.3}$$

(3) **Variances:**

$$\bar{\sigma_i^2} = \frac{\sum_{t=1}^{T} P_r(i/x_t, \lambda) x_t^2}{\sum_{t=1}^{T} P_r(i/x_t, \lambda)} - \bar{\mu_i^2}. \tag{3.4}$$

GMM learns features of genuine and spoofed speech from the given training speech signals and generates a statistical model. In the testing (evaluation) phase, the SSD system analyze the incoming utterance and then estimates the Log-likelihood Ratio (LLR) using pre-trained GMM parameters.

3.3.2 Convolutional Neural Network (CNN)

CNN is a neural network-based architecture, which consist of one or more convolutional layers followed by classification layers [66]. In this work, the input feature size for CNN is taken to be 30×400. Our CNN architecture consist of five

convolutional layers (Conv1, Conv2, Conv3, Conv4, and Conv5) followed by two fully-connected layers (FC1 and FC2). Here, in the first two convolutional layers, the data is convolved using a kernel size of 5×5 with a stride of 1 and padding of 2. Furthermore, in the remaining three convolutional layers, the kernel is used of size 3×3 with the stride and padding of 1. Here, after every convolutional layer, we have used max-pool layer having kernel of size 2×2 with a stride of 2, in order to reduce the size of data and also to reduce the computation cost of CNN model. After extraction of features from the convolutional layers, the output of Conv5 is fed to FC1 layer and the probabilistic output for classification is taken from FC2. The Rectified Linear activation Unit (ReLU) function is taken as the activation function for all hidden as well as FC layers [67]. Binary cross-entropy is taken to be the loss function and for optimization of weights, we have used stochastic gradient descent method [68].

3.3.3 Light-CNN (LCNN)

LCNN is modified version of CNN, which consist CNN with Max-Feature-Map (MFM) activation function. It is defined as [51]:

$$\begin{aligned} y_{ij}^{k} &= max(x_{ij}^{k}, x_{ij}^{k+\frac{N}{2}}), \\ \forall i &= \overline{1,H}, j = \overline{1,W}, k = \overline{1,N/2}. \end{aligned} \tag{3.5}$$

Here, x is the input feature vector of size $H \times W \times N$, and y is the output feature vector of size $H \times W \times N/2$. Furthermore, i and j is represents frequency and time-domain, respectively, and the value of k indicates channel index.

For our experiments, we have used input feature of size 30×400 for LCNN model. The LCNN model consists of four CNN layers (Conv1, Conv2, Conv3, and Conv4) and two FC layers (FC1, FC2). In the convolutional layers, the data are convolved using a kernel of size 3×3 with a stride of 1 and padding of 1. After each layer, the MFM and max-pooling layer is used. The MFM layer uses a kernel of size 3×3 with stride of 1 and padding of 2. The max-pooling is used with kernel size of 2×2 and stride of 2, to reduce the size of feature vector and also to reduce the complexity of the model. The ReLU activation function is used in FC7 layer to discriminate between genuine and spoofed class. For calculation of loss, we have used binary cross-entropy as loss function and for optimization of weights, we have used stochastic gradient descent method.

3.4 Classifier Used for VLD

3.4.1 CNN

A Convolution Neural Network (CNN) or ConvNet [69, 70] is a neural network model that consists of one or more convolutional layers followed by a classification layer. The two wavelet-based approaches described in sub-Section 6.3, which yield matrices of sizes 45×45 and $3 \times 512 \times 512$, respectively. For our experiments, the CNN architecture (shown in Figure 3.1) consists of 3 convolutional layers (Conv1, Conv2, Conv3) followed by 3 Fully-Connected (FC1, FC2, and FC3). The output of Conv3 is fed to the FC1 layer. The output of the final FC3 layer provides a probabilistic output for classification. Sigmoid activation function used at the output of FC3, while ReLU activation function is used for all the hidden layers. Binary cross-entropy is used as the loss function and stochastic gradient descent algorithm is used as the optimization algorithm. The sequence and the number of layers in the CNN are kept the same for *45*-D handcrafted features as well as scalogram. However, for the case of scalogram, images of size 512×512, the input is convolved with a kernel of size 7×7 for Conv1 and 3×3 for Conv2 and Conv3. For the case of handcrafted *45*-D wavelet-based features, the input is convolved with a kernel of size 3×3 during the forward pass, with a stride of 1, and zero-padding of 1. A max-pooling layer with a kernel size of 3×3, and stride of 1 is used.

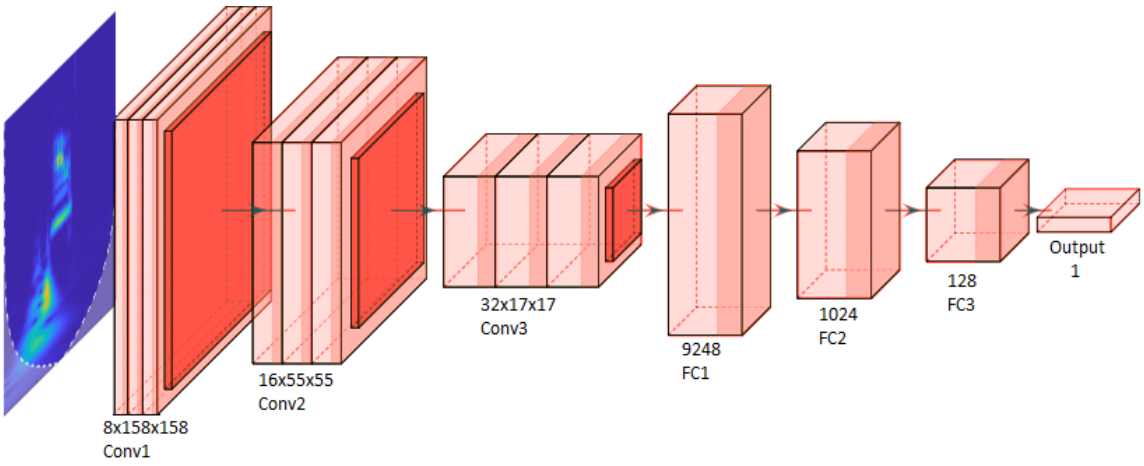


Figure 3.1: The CNN architecture used for classification of the proposed Morlet wavelet scalogram-based features. After [8].

3.5 Performance Metrics

The performance metrics used in this work are % Equal Error Rate (EER) and % classification accuracy. As we discussed in the sub-Section 3.3.1, the LLR scores are estimated for testing data using a pretrained GMM. The LLR scores are used to compute False Rejection Ratio (FRR) and False Acceptance Ratio (FAR). Hence, the value of EER is, where FRR equals to FAR. Hence, the % EER is given by:

$$\%EER = \frac{FAR + FRR}{2} \times 100. \tag{3.6}$$

While plotting a Detection Error Trade-off (DET) curve, we plot error rates on both the axes, giving uniform treatment to both the types of errors, and use a scale for both axes which, spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear.

For calculation of % classification accuracy, first step is to use a classification model, which make a prediction of class labels for each sample of testing dataset. The predicted labels are then compared with actual labels of testing data. The % classification accuracy is then calculated based on the correct prediction of classification model. The prediction of labels by classification model is divided in four parts, namely, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The % classification accuracy is calculated from these four parts as [71]:

$$\%\text{ Classification Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \tag{3.7}$$

3.6 Data Fusion Strategies

In our experiments, we have used score-level fusion technique on LLR scores, which is evaluated from the multiple SSD systems. With the help of this score-level fusion, we can capture the possible complementary information from different SSD systems. The calculation of score-level fusion for two feature sets using *linear weighted sum* is given by:

$$LLR_{fused} = \alpha_i \cdot LLR_{feature1} + (1 - \alpha_i) \cdot LLR_{feature2}, \tag{3.8}$$

where $LLR_{feature1}$ and $LLR_{feature2}$ are the LLR scores calculated from the feature set-1 (system 1) and feature set-2 (system 2), respectively. The fusion parameter α_i and $(1 - \alpha_i) \in [0, 1]$ show the contribution of the individual systems during

score-level fusion.

3.7 Chapter Summary

This chapter explains the details of dataset used in this study, such as ASVSpoof 2017 v2.0, BTAS 2016, ReMASC, and POCO. Furthermore, the detailed architecture of classifiers are explained which are used in various tasks. Later, the evaluation strategies used in this work are explained along with the data fusion technique. In the next chapter, the Quadrature Energy Separation Algorithm-based feature is explained for replay SSD.

CHAPTER 4
Instantaneous Frequency Estimation Using Quadrature Energy Separation Algorithm

4.1 Introduction

In a recent study, the TEO is used to estimate Instantaneous Frequency (IF) via Energy Separation Algorithm (ESA). Hence, Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency via ESA (CFCCIF-ESA) was proposed, which utilizes only the amplitude information of the signal for replay SSD. Moreover, due to absence of Hilbert-transform (HT), it does not contain the quadrature-phase component of the signal. Therefore, in order to incorporate both the advantages, i.e., excellent time resolution of TEO and having quadrature-phase component via HT, we propose CFCCIF-QESA feature set. Here, the term QESA represents Quadrature-based ESA. Furthermore, QESA is based on the extended definition of TEO for complex signals. To our knowledge, this extended definition of TEO is exploited for the first time for SSD task. Additionally, the choice of quadrature-phase (90°) component along with in-phase component is justified by Mutual Information (MI)-based analysis, described in further detail in Section 4.3. As a result, we have developed CFCCIF-QESA feature set [1].

4.2 Estimation of Instantaneous Frequency (IF)

This Section describes the conventional methodology for IF estimation. In particular, two methods of IF estimation is shown, such as (i) IF Estimation Using Analytic Signal, and (ii) IF Estimation Using ESA.

[1] This work is jointly done with PhD. scholar at DA-IICT, Ms. Priyanka Gupta.

4.2.1 IF Estimation Using Analytic Signal

The IF of a real signal is defined as the time derivative of the unwrapped phase of the analytic signal, whose Fourier transform is zero for negative frequencies. The analytic signal $x_a(t)$ corresponding to a real signal $x(t)$ is given by:

$$x_a(t) = x(t) + j\hat{x}(t), \tag{4.1}$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$. The corresponding analytic (or instantaneous) phase ϕ(t) and IF are given by:

$$\phi(t) = arctan\left(\frac{\hat{x}(t)}{x(t)}\right), \tag{4.2}$$

$$IF = \frac{d(\phi(t))}{dt}. \tag{4.3}$$

The use of arctangent function in eq.(4.2) creates a signal processing artefact (due to the periodicity property of arctan) called as phase wrapping, thereby creating discontinuities in the phase function, $\phi(t)$. Due to this discontinuity, the IF cannot be derived directly from $\phi(t)$ using eq. (4.3) without the computationally complex task of phase unwrapping [72].

4.2.2 IF Estimation Using ESA

The TEO $\Psi\{.\}$ of a continuous-time real signal $x(t)$ is defined as [73]:

$$\Psi\{x(t)\} = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \tag{4.4}$$

where $\dot{x}(t)$ denotes the first-order derivative of $x(t)$, and $\ddot{x}(t)$ denotes the second-order derivative of $x(t)$ w.r.t. time t. Furthermore, for a discrete-time signal $x[n]$, the TEO is defined mathematically approximating the derivative operation in eq. (4.4) [73]. In particular,

$$\Psi\{x[n]\} = x^2[n] - x[n-1]x[n+1]. \tag{4.5}$$

TEO tracks rapid energy (or its running estimate) of the speech signal *within* a glottal cycle with excellent time resolution, requiring only three consecutive samples [49, 73]. Moreover, the TEO enables to estimate the Amplitude Modulation (AM) and Frequency Modulation (FM) components of a speech signal, by the well known ESA which is described next.

The time-varying amplitude and frequency behaviour in a speech signal is modelled as an AM-FM signal [74]. In particular,

$$\begin{aligned} x[n] &= a[n]cos[\phi[n]], \\ &= a[n]cos\left[\omega_c n + \omega_m \int_0^n q(m)d(m) + \theta\right], \end{aligned} \tag{4.6}$$

where the maximum deviation in frequency is $|q[n]| \leq 1$, $\omega_m \in [0, \omega_c]$, $a(n)$ is instantaneous amplitude, and θ is the constant offset. The instantaneous frequency $\omega[n]$ is given by [75]:

$$\omega[n] = \frac{d}{dt}\phi[n] = \omega_c + \omega_m q[n], \tag{4.7}$$

where ω_c is the carrier frequency. Furthermore, TEO applied on AM-FM signals (such as shown in eq. (4.6)), approximately estimates the product of instantaneous amplitude and instantaneous frequency [74,76]. In particular,

$$\Psi\left[a[n]cos\left[\int_0^n \Omega[m]dm + \theta\right]\right] \approx a^2[n]sin^2(\omega[n]) = a^2[n].\omega^2[n], \tag{4.8}$$

where $sin^2(\omega[n]) \approx \omega^2[n]$, for $\omega << \omega_c$. Thus, it can be observed that both $a[n]$ and $w[n]$ contributes to running estimate of energy of AM-FM signal representing Simple Harmonic Motion (SHM) [46]. Hence, the following expressions for $a[n]$ and $\omega[n]$ are called as Energy Separation Algorithm (ESA) [46]:

$$a[n] \approx \frac{2\Psi(x[n])}{\sqrt{\Psi(x[n+1] - x[n-1])}}, \tag{4.9}$$

$$\omega[n] \approx arcsin\left(\sqrt{\frac{\Psi(x[n+1] - x[n-1])}{4\Psi(x[n])}}\right). \tag{4.10}$$

4.3 Exploiting Relative Phase-Based Information

So far, most of the features have been derived from the magnitude spectrum of the speech signal [77]. However, the phase characteristics can also be useful for many applications [78–81]. In this work, we employ an information-theoretic, approach to measure *relative* phase-based information, without estimating phase explicitly. In particular, we use Mutual Information (MI) to analyze the amount of information between the signal and its corresponding phase-shifted signal [82]. MI of two signals is a measure of dependence of the signals on each other, i.e., a measure of

how much information the two signals share. It tells us how much knowing one of the two signals reduces uncertainty about the other. For example, if two signals X and Y are independent, then knowing X does not yield any information about Y and vice-versa, so their MI is zero. Mathematically, MI is estimated as [82]:

$$I(X;Y) = h(X) - h(X|Y), \tag{4.11}$$

where h denotes the entropy (i.e., measure of randomness). Using the joint and marginal Probability Density Function of X and Y, the MI is [82]:

$$I(X;Y) = \int_x \int_y f_{XY}(x,y) \log_2 \left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right) dydx. \tag{4.12}$$

Given that the speech signal can be modelled as an AM-FM signal, we consider an AM-FM signal as:

$$\begin{aligned} a(t) &= (1 + 0.5cos\,(60\pi t)), \\ x(t) &= a(t) \cos\left(2\pi f_c t + 4sin\left(2\pi f_c t + \left(\frac{\pi}{4}\right)\right)\right). \end{aligned} \tag{4.13}$$

For this AM-FM signal (expressed via eq. (4.13)) and its phase-shifted version, we have estimated the MI. The angle at which MI is minimum is the *optimum* phase value. From the values of MI obtained (as shown in Fig. 4.1), it can be observed that the optimum phase difference is 90° with MI=1.4349 bits. In addition, for a signal $x(t)$, the Fourier transform is denoted as $X(\omega) = X_R(\omega) + jX_I(\omega)$. Therefore, $X(\omega)$ is given as

$$X(\omega) = tan^{-1}\left(\frac{X_I(\omega)}{X_R(\omega)}\right). \tag{4.14}$$

From eq. (4.14), it can be observed that the Fourier transform phase $X(\omega)$ is always zero for $X_I(\omega) = 0$, which means if we do not use $\pi/2$-shifted version of $\cos(\omega t)$ (i.e., $\sin(\omega t)$) as an additional basis function in the definition of Fourier transform; it is not possible to compute $X(\omega)$. In this context, Fig. 4.1 (b) shows the MI obtained between a cosine and its phase-shifted versions. Notably, for the cosine signal as well, MI is observed to be minimum at $\pi/2$ phase shift in $cos(\omega t)$ (i.e., $\sin(\omega t)$) indicating significance of $\cos(\omega t)$ (i.e., in phase) and its quadrature component (i.e., $\sin(\omega t)$) in the original definition of the Fourier transform.

To that effect, taking phase shift as 90° (i.e., a quadrature), we propose an improved relative phase-based CFCCIF-QESA feature set. The feature extraction procedure of CFCCIF-QESA is shown in Algorithm 1. The quadrature component of the real-valued speech signal is achieved using Hilbert transform, which results

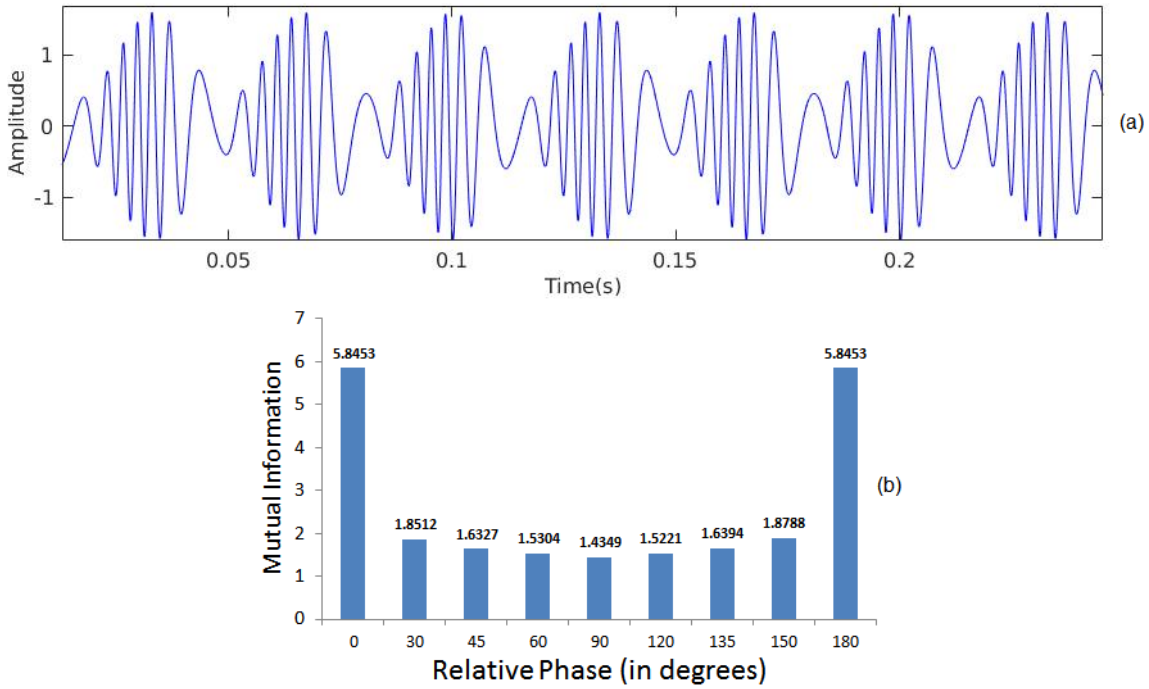


Figure 4.1: (a) AM-FM signal, and (b) MI between AM-FM signal and its phase-shifted version.

in a complex-valued analytic signal, having a *causal* spectrum. Subsequently, TEO for complex signals is used for estimating IF using ESA. In the next sub-Section, we present the extended definition of TEO for complex signals, which is further used in the CFCCIF-QESA feature extraction procedure.

4.4 Extracting TEO-Based Energy for Complex Signals

As discussed above, we exploit quadrature phase-shift by estimating analytic signal. Here, we discuss the extended definition of the TEO on a complex-valued signal $z(t)$, i.e., $\psi_c[z(t)]$ which is given by [83]:

$$\psi_c[z(t)] = z(t)\dot{z}^*(t) - \frac{1}{2}[\ddot{z}(t)z^*(t) + z(t)\ddot{z}^*(t)]. \tag{4.15}$$

Given that $z(t)$ is complex, the TEO defined in eq. (4.15) is applied on real and imaginary parts of $z(t)$ as

$$\psi_c[x_z(t)] = \dot{x_{z_R}}^2(t) + \dot{x_{z_I}}^2(t) - x_{z_R}(t)\ddot{x}_{zR}(t) - x_{z_I}(t)\ddot{x}_{zI}(t). \tag{4.16}$$

When $z(t)$ is complex, eq. (4.16) can be re-written as [84]:

$$\psi_c[z(t)] = \psi[z_r(t)] + \psi[z_i(t)]. \tag{4.17}$$

In this work, we extract TEO-based energy using eq. (4.17) on complex-valued analytic signal for improved estimation of energy as a part of ESA, discussed in the next sub-Section.

4.5 CFCCIF-QESA Feature Extraction

The proposed CFCCIF-QESA feature set consists of various sub-systems, as shown in Fig. 4.2. The filterbank of the CFCCIF-QESA consists of AT-based cochlear filters, which represent the human auditory system consisting of Basilar Membrane (BM). As per place theory of hearing [75], only a particular region of the BM vibrate in response to a particular frequency region in the speech signal. The inner hair cells act as transducers, converting the vibrations of the BM to energy. Given that the motion of the hair cell is only in the *positive* direction, it is expressed mathematically as:

$$H(a,b) = (F(a,b))^2, \tag{4.18}$$

where $F(a,b)$ is the output of the filterbank, and a and b govern the *size* and *shape* of each cochlear filter. The hair cell output of each filterbank is converted into a representation of the nerve spike density, which is computed as an average of $H(a,b)$ [31]. Furthermore, the quadrature-phase component in the output $f[n]$ of

Algorithm 1: IF estimation using Quadrature-based Energy Separation Algorithm (QESA)

Input: Subband filter output $f[n]$ Output: IF $f_z[n] = f[n] + j.HT\{f[n]\}$
`/* Using Equation(4.17) */`
$E_r[n] \leftarrow \text{TEO}\{\text{real}(f_z[n])\}$
$E_i[n] \leftarrow \text{TEO}\{\text{imag}(f_z[n])\}$
$\psi\{f_z[n]\} = E_r[n] + E_i[n]$
$IF \leftarrow Cos^{-1}\left[\frac{1-\psi\{f_z[n]-f_z[n-1]\}}{2\psi\{f_z[n]\}}\right]$

the filterbank is introduced by taking its analytic signal, $f_z[n]$. This is because the analytic signal is generated by taking the Hilbert transform, which is nothing but the quadrature-shifted version of $f[n]$. Now, in order to estimate the energy of the complex-valued analytic signal, we use the extended definition of TEO as described in Section 4.4. Furthermore, the energy profile obtained from the extended

TEO is used to estimate IF using ESA. The ESA for IF estimation of a signal $f(n)$ is given by [75]:

$$\omega_i[n] \approx Cos^{-1}\Big[\frac{1-\psi\{f[n]-f[n-1]\}}{2\psi\{f[n]\}}\Big]. \tag{4.19}$$

Here, the $\psi\{f[n]\}$ represents the Teager energy of $f[n]$, and the $\omega_i[n]$ represents the estimated IF.

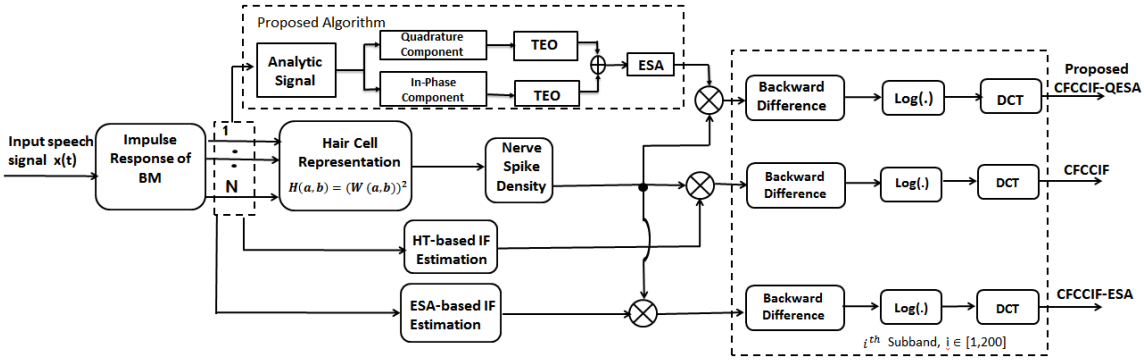


Figure 4.2: Functional block diagram of the proposed CFCCIF-QESA feature set, along with conventional CFCCIF and CFCCIF-ESA feature sets. After [6].

4.6 Spectrographic Analysis of CFCCIF-ESA *vs.* CFCCIF-QESA

To analyse the effectiveness of CFCCIF-QESA as compared to the CFCCIF-ESA, we observe the spectrograms of both the feature sets for the same speech utterance, which is shown in Fig. 4.3. Here, the spectrogram of a feature set represents the Energy Spectral Density (ESD) in the time-frequency plane for that feature set. However, this spectral representation is obtained before DCT operation in the feature extraction. In Fig.4.3, Panel I and Panel II show the ESD for CFCCIF-ESA and CFCCIF-QESA, respectively. Fig. 4.3(a) and Fig. 4.3(b) represent a genuine speech utterance and corresponding spoofed (replay) speech utterance, respectively. It can be observed from Fig. 4.3 that in the higher frequency region, the resolution of CFCCIF-QESA feature set is high for both genuine and spoofed as compared to the CFCCIF-ESA feature set. It is highlighted in Fig. 4.3 by curved rectangle boxes. Furthermore, the oval boxes show the difference in ESD of genuine *vs.* spoofed speech, for CFCCIF-QESA feature set. Thus, we believe that due to the additional information of quadrature phase in CFCCIF-QESA, we can get high ESD for both genuine and replayed utterances because all the other computational details are controlled in both the feature sets except the quadrature phase

component.

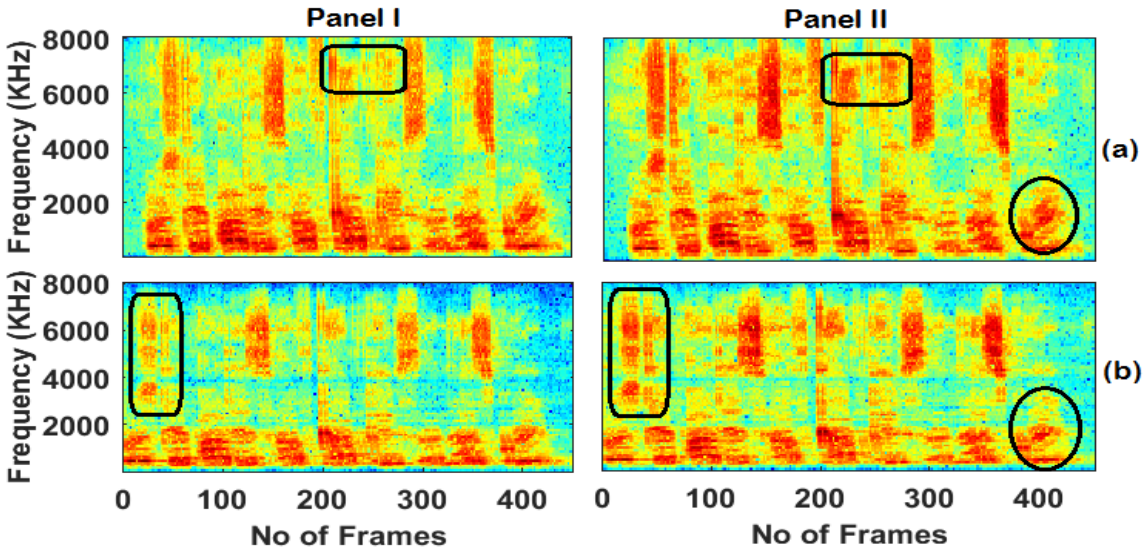


Figure 4.3: Spectrographic representation of the genuine *vs.* spoofed speech. Panel I and Panel II represent spectrographic representation of CFCCIF-ESA and CFCCIF-QESA, respectively. Here, (a) genuine speech signal, and (b) corresponding spoofed (replay) speech signal. After [6]

4.7 Experimental Results

This Section shows the experimental results obtained using the proposed feature set along with existing features for replay SSD using ASVSpoof 2017 v2.0 and BTAS 2016 dataset.

4.7.1 Results on ASVSpoof 2017 v2.0 Database w.r.t. Various Classifiers

The results of CFCCIF-QESA are compared with the other existing feature sets using GMM, CNN, and LCNN, as shown in Table 4.1, Table 4.2, and Table 4.3, respectively.

It can be observed that the proposed feature set (denoted by S5) performs the best as compared to the other systems (i.e., S1 to S4). To be specific, we achieve EER of 11.40% and accuracy of 73.35% on the ASVSpoof 2017 evaluation set. To emphasize the benefit of incorporating quadrature phase component, the results show that the proposed system S5 (i.e., with quadrature phase component) gives an absolute decrease in EER of 3.37% and an absolute improvement of 4.83% in

Table 4.1: Results on ASVSpoof 2017 v2.0 database using GMM. After [6].

Feature Set	Dev.		Eval.	
	% EER	% Accu.	% EER	% Accu.
CQCC (S1)	12.87	81.75	18.81	59.72
CFCC (S2)	17.60	79.29	18.97	59.96
CFCCIF (S3)	16.61	78.59	17.38	58.13
CFCCIF-ESA (S4)	11.54	82.57	14.77	68.52
CFCCIF-QESA (S5)	**9.48**	**87.30**	**11.40**	**73.35**
S1+S5	9.48	87.30	11.40	73.35
S2+S5	9.47	87.34	11.39	73.36
S3+S5	9.37	87.34	11.38	73.39
S4+S5	9.25	87.70	11.31	73.80
S2+S3+S4+S5	9.22	87.90	11.25	73.96
S1+S2+S3+S4+S5	**9.21**	**87.94**	**11.24**	**74.03**
+ indicates score-level fusion as per eq. (3.8)				

accuracy, as compared to system S4 (with no quadrature phase component). Furthermore, we performed score-level fusion as per eq. (3.8) (denoted by + in Table 4.1) of system S5 with all the remaining systems S1 to S4. The score-level fusion of three systems, which are based on cochlear filtering (i.e., S3, S4, and S5) further reduced the EER to 9.36% and 11.19% on the development and evaluation sets, respectively.

Table 4.2: Results on ASVSpoof 2017 v2.0 database using CNN. After [6].

Feature Set	Dev.		Eval.	
	% EER	% Accu.	% EER	% Accu.
CQCC (S1)	5.38	93.56	20.77	55.23
CFCC (S2)	5.06	94.26	21.45	54.10
CFCCIF (S3)	12.92	86.90	20.53	55.70
CFCCIF-ESA (S4)	13.92	85.02	19.26	56.34
CFCCIF-QESA (S5)	**9.74**	**88.36**	**19.10**	**57.40**
S1+S5	2.36	97.48	12.87	71.45
S2+S5	7.30	92.32	17.90	58.10
S3+S5	8.77	90.64	17.52	58.17
S4+S5	9.19	88.77	17.27	58.80
S2+S3+S4+S5	7.10	92.88	16.45	59.30
S1+S2+S3+S4+S5	**1.88**	**97.60**	**12.45**	**72.10**

Table 4.2 shows the performance when CNN was used as the classifier. The proposed feature S5 achieves better performance as compared to the cochlear filter-based features (i.e, S2, S3, and S4). An absolute decrease in EER of 0.16%, and an absolute improvement of 1.06% in accuracy, is observed as compared with

system S4. It should be noted that even though this absolute improvement is not very significant, we achieve EER of 12.45% and accuracy of 72.10%, when S5 is fused with all the remaining feature sets.

Table 4.3: Results on ASVSpoof 2017 v2.0 database using LCNN. After [6].

Feature Set	**Dev.**		**Eval.**	
	% EER	**% Accu.**	**% EER**	**% Accu.**
CQCC (S1)	7.00	90.40	30.11	40.21
CFCC (S2)	5.92	93.45	26.47	51.30
CFCCIF (S3)	13.36	85.61	20.29	55.50
CFCCIF-ESA (S4)	13.08	86.43	18.05	58.20
CFCCIF-QESA (S5)	**11.22**	**87.10**	**17.52**	**59.30**
S1+S5	3.51	95.08	15.00	63.10
S2+S5	3.84	95.96	15.22	62.63
S3+S5	9.83	89.88	16.49	61.10
S4+S5	9.28	90.05	15.96	61.45
S2+S3+S4+S5	2.31	97.60	14.30	65.01
S1+S2+S3+S4+S5	**2.29**	**97.71**	**13.71**	**67.30**

Table 4.3 shows the performance when LCNN was used as the classifier. We observe better performance of S5 with LCNN as compared to the CNN. In particular, we obtain an EER of 17.52% and an accuracy of 59.30% on the evaluation set of ASVSpoof 2017 database. Furthermore, performance behaviour similar to GMM and CNN can be observed as S5 performs better when compared to all the cochlear filter-based features (i.e., S2, S3, and S4). This also confirms the significance of quadrature phase component in the proposed feature set. Furthermore, if we compare the performance of individual feature sets (from S1 to S4), with their individual fusion performance with S5 (i.e., S1+S5, S2+S5, S3+S5, and S4+S5), we observe improvement in performance for *each* fusion case. To that effect, on the evaluation set of ASVSpoof 2017, the maximum absolute decrease in EER of 15.11% and 22.89 in accuracy is observed w.r.t. S1 and S1+S5, as shown in the Table 4.3.

Classifier-Level Fusion: Given various classifiers (i.e., GMM, CNN, and LCNN) were used on the ASVspoof 2017 v2.0 dataset, we now present the classifier-level fusion results in Table 4.4. It shows the results obtained on the proposed CFCCIF-QESA using GMM, CNN, and LCCN, labelled as S1, S2, and S3, respectively. The best performance on the evaluation set is observed when the scores of all the three classifiers are fused, leading to an EER of 11% and an accuracy of 74.02%. Notably, CFCCIF-QESA shows relatively the best performance using GMM. The better per-

formance of GMM can be due to the data being more approximated to Gaussian and the characteristics are better suited for GMM. Notably, in the ASVspoof 2021 PA challenge, the LFCC-GMM baseline (with 39.79% ERR) showed better performance as compared to the LFCC-LCNN baseline (with 42.16% ERR). This also shows that based on distributional characteristics of data, GMM can indeed perform better than the neural network-based classifiers, such as CNN and LCNN.

Table 4.4: Results of classifier-level fusion of CFCCIF-QESA feature set using different classifiers on ASVSpoof 2017 v2.0 dataset. After [6].

Classifier Used	**Dev.**		**Eval.**	
	% EER	**% Accu.**	**% EER**	**% Accu.**
GMM (S1)	9.48	87.30	11.40	73.35
CNN (S2)	9.74	88.36	19.10	57.40
LCNN (S3)	11.22	87.10	17.52	59.30
S1+S2	6.97	90.00	11.40	73.35
S1+S3	7.79	89.82	11.00	74.00
S2+S3	8.78	89.75	16.55	66.36
S1+S2+S3	**6.62**	**90.99**	**11.00**	**74.02**

4.7.2 Results on BTAS 2016 Dataset

The BTAS 2016 dataset is an extended version of ASVSpoof 2015 dataset. In particular, it contains VC, SS, and replay spoofed utterances. The experimental results using proposed feature set and the other features are shown in Table 4.5. It can be noted that CFCCIF-QESA performs relatively close to the CFCCIF-ESA feature set. However, we observe relatively the best performance in EER when all the features are fused to give an EER of 3.43% and an accuracy of 93.67%.

4.8 Chapter Summary

In this chapter, auditory transform-based CFCCIF-QESA feature set is proposed. MI-based analysis is done to determine the optimum relative phase shift. It is found that a quadrature phase shift is the best suited. Further, MI is to justify basis functions used in the original definition of Fourier transform. To that effect, the signal is converted to its analytic signal (which has its real and imaginary parts separated by a quadrature phase). The analytic signal is complex-valued and hence, for the first time, the extended definition of TEO for complex signals is used for the SSD task. Experiments are performed on ASVspoof 2017 version 2.0

Table 4.5: Results (in % EER and % classification Accuracy) on BTAS 2016 dataset using GMM. After [6].

Feature Set	**Dev.**		**Eval.**	
	% EER	**% Accu.**	**% EER**	**% Accu.**
CQCC (S1)	2.57	91.50	4.45	88.32
CFCC (S2)	1.98	92.61	4.18	90.08
CFCCIF (S3)	2.13	92.00	7.35	81.13
CFCCIF-ESA (S4)	2.07	92.11	5.02	86.23
CFCCIF-QESA (S5)	**1.81**	**93.00**	**5.20**	**86.00**
S1+S5	1.81	93.00	3.90	91.20
S2+S5	1.77	93.32	4.01	91.70
S3+S5	1.81	93.00	5.20	86.00
S4+S5	1.81	93.01	5.01	86.25
S2+S3+S4+S5	1.71	93.88	3.85	92.33
S1+S2+S3+S4+S5	**1.63**	**94.23**	**3.43**	**93.67**

and BTAS 2016 datasets and CFCCIF-QESA features are shown to perform better than features without quadrature-phase on ASVSpoof 2017 version 2.0 using GMM, CNN, and LCNN. Furthermore, the similar behaviour of proposed feature set can be observed for BTAS 2016 dataset using GMM. The future research efforts regarding this study, will be directed towards investigating the significance of the proposed feature set on the other spoofing attacks, such as VC and SS on ASVSpoof 2015 challenge dataset and on the recently released DeepFake speech data of ASVSpoof 2021 challenge. In addition, apart from CNN and LCNN used in this work, we plan to investigate the other deep learning-based classifiers, such as ResNet and LSTM. In the next chapter, we present the relevance of effective beamforming from the perspective of replay SSD on VAs.

CHAPTER 5
Significance of Beamforming Technique for Replay SSD

5.1 Introduction

This chapter investigates the capability of the Delay and Sum (DAS) beamformer to extract the reverberation characteristics in replay speech signals. The replay mechanism consists of the characteristics of the recording, playback devices, and corresponding environments due to which reverberation characteristics are embedded into the replay speech signal. Further, analysis is presented for DAS *vs.* MVDR beamformer for replay SSD task. MVDR is a state-of-the-art beamformer for speech enhancement applications, as it successfully *nullify* the reverberation effects in distant speech signals. Whereas, DAS suppresses the additive noise and retains the reverberation effect observed in the output signal and hence, DAS is suitable choice for replay SSD task.

5.2 Speech Signal Modeling for Microphone Array Signal

Assuming the linear and time-invariant (LTI) model for the acoustic medium (path) between speech source and microphone array, the speech signal received by N-element microphone array is given as [75, 85–87]:

$$\begin{aligned} x_i(n) &= r_i(n) * k(n) + \eta_i(n), \\ &= y_i(n) + \eta_i(n), i = 1, 2, ..., N, \end{aligned} \tag{5.1}$$

where i represents the index for i^{th} microphone in array, $r_i(n)$ is the impulse response of the acoustic medium between desired source signal $k(n)$ and i^{th} microphone. '*' represents the convolution operation and $\eta_i(n)$ corresponds to additive

noise of the i^{th} microphone. Here, for modelling of noisy speech signal $x_i(n)$, it is assumed that the speech signal $y_i(n)$ and noise signal $\eta_i(n)$ are zero-mean and uncorrelated. During development of the replay speech signal, impulse responses of recording devices ($b(n)$) and environment ($c(n)$) as well as impulse responses of playback devices ($e(n)$) and environment ($f(n)$) are convolved with the source signal. Let $a(n)$ represents the combination of these impulse responses [88], i.e., :

$$a(n) = b(n) * c(n) * e(n) * f(n). \tag{5.2}$$

Hence, the replay speech signal ($x_{ir}(n)$) can be represented as:

$$\begin{aligned} x_{ir}(n) &= r_i(n) * a(n) * k(n) + \eta_i(n), \\ &= y_{ir}(n) + \eta_i(n), i = 1, 2, ..., N. \end{aligned} \tag{5.3}$$

Thus, the characteristics of the $y_{ir}(n)$ in eq. (5.3) is different from that of $y_i(n)$ because of the additional impulse response $a(n)$ caused by the replay mechanism. Considering this $a(n)$ as distinguishing acoustic characteristics of the replay spoof, it can be emphasized using suitable signal processing technique for replay SSD. To that effect, first we present the significance of the DAS beamformer over MVDR for replay SSD through mathematical analysis, and then it is validated using suitable experiments.

The representation of the received signal in eq. (5.1) in frequency-domain can be expressed as [86]:

$$\begin{aligned} X_i(\omega) &= R_i(\omega) \odot K(\omega) + H_i(\omega), \\ &= Y_i(\omega) + H_i(\omega), \quad i = 1, 2, ..., N, \end{aligned} \tag{5.4}$$

where $X_i(\omega)$, $R_i(\omega)$, $K(\omega)$, $H_i(\omega)$, and $Y_i(\omega)$ are the discrete-time Fourier transforms (DTFTs) of $x_i(n)$, $r_i(n)$, $k(n)$, $\eta_i(n)$, and $y_i(n)$, respectively. Here, the symbol $\odot$ represents the componentwise multiplication operation (due to convolution theorem for Fourier transform). The frequency-domain representation of N-microphone array can be represented in the matrix form as :

$$\mathbf{X}(\omega) = \mathbf{R}(\omega) \odot K(\omega) + \mathbf{H}(\omega) = \mathbf{Y}(\omega) + \mathbf{H}(\omega), \tag{5.5}$$

where

$$\begin{aligned}
&\mathbf{X}(\omega) = [X_1(\omega), .., X_N(\omega)]^T, \mathbf{R}(\omega) = [R_1(\omega), .., R_N(\omega)]^T, \\
&\mathbf{K}(\omega) = [K(\omega), .., K(\omega)]^T, \mathbf{Y}(\omega) = [Y_1(\omega), .., Y_N(\omega)]^T, \\
&\mathbf{H}(\omega) = [H_1(\omega), .., H_N(\omega)]^T.
\end{aligned} \tag{5.6}$$

5.3 Beamforming

5.3.1 Delay and Sum (DAS) Beamformer

The DAS is a primitive beamforming technique for noise reduction in the array signal processing literature [89, 90]. This involves reinforcing the desired signal while suppressing the unwanted noise signals. The conventional DAS beamformer will delay all the input signals in time *w.r.t.* the reference signal, such that the array sensor can focus in one direction. Hence, the summation of the delayed signals with the reference signal will result in suppression of noise, which is arriving from the other directions. Furthermore, it can be postulated that the summation of the delayed signals leads to cancellation of *additive (random) noise*. Figure 5.1 shows the functional block diagram of DAS beamformer from receiver end. Here, weights for corresponding single channel microphone signal in a microphone array are shown. The time-domain representation of DAS beamformer is given by [91]:

$$\mathbf{d}(n) = \frac{1}{\beta} \sum_{i=1}^{N} w_i x_i(n - \tau_i). \tag{5.7}$$

Furthermore, the frequency-domain representation of DAS beamformer is given by taking DTFT of eq. (5.7) [92]. In particular,

$$\mathbf{D}(\omega) = \frac{1}{\beta} \sum_{i=1}^{N} w_i X_i(\omega) e^{-j\omega\tau_i} = \mathbf{W}^H \mathbf{X}(\omega), \tag{5.8}$$

$$\text{where} \quad \mathbf{W} = \frac{1}{\beta} \sum_{i=1}^{N} w_i e^{-j\omega\tau_i}, \tag{5.9}$$

where w_i is the elementwise weighting for the spatial window, β is the summation of the weights, and $\mathbf{W}$ is the steering vector (optimized weight vector) of desired *linear* phase shift and weights. The superscript H denote the Hermitian transpose. In fact, it should be noted that it is due to this linear phase filtering, acoustic char-

acteristics of replay are preserved in DAS beamformed signal. The $\mathbf{D}(\omega)$ represents the frequency response of beamformed signal. In the framework of Wiener-Khinchin theorem, the power at the output of the beamformer is estimated by taking the Fourier transform of the autocorrelation function of the beamformer output [93], i.e.,

$$\mathbf{p}(\omega) = \mathbb{E}[|\mathbf{D}(\omega)|^2], \tag{5.10}$$

where $\mathbb{E}[\cdot]$ is the expectation operator.

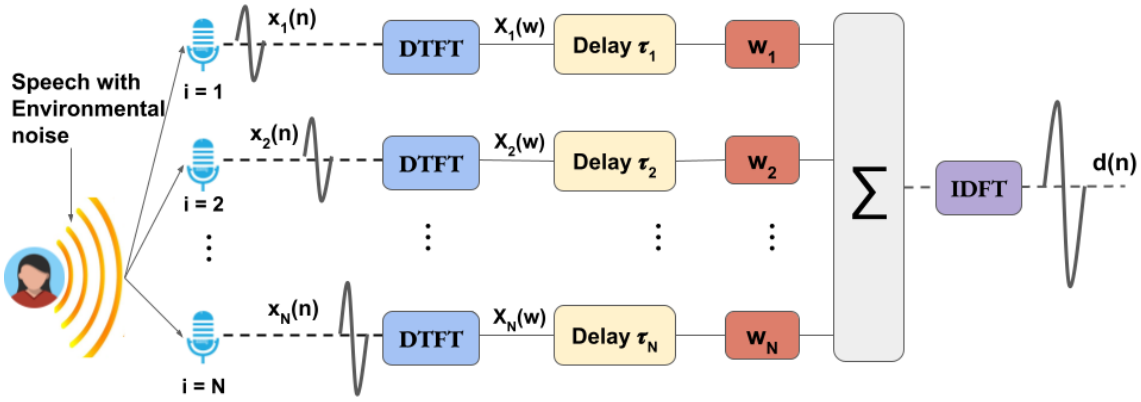


Figure 5.1: Functional block diagram of DAS beamformer having N the number of microphones in array. After [9].

5.3.2 Minimum Variance Distortionless Response (MVDR)

MVDR beamformer achieves the speech enhancement by suppressing (ideally nullifying) the reverberation effects introduced by the room acoustics [85, 94]. In this approach, Signal-to-Noise Ratio (SNR) of the multi-channel audio signal is significantly improved by minimizing the distortion (noise) [95]. For this formulation, it is assumed that the audio signal from the reference source is distortionless, which also results in preservation of all-pass characteristics. However, MVDR increases the computational complexity of the system. The matrix for output power $\mathbf{p}(\omega)$ of MVDR beamformer is given by:

$$\mathbf{p}(\omega) = \mathbb{E}[|\mathbf{D}(\omega)|^2] = \mathbf{W}^H \mathbf{V}(\omega) \mathbf{W}, \tag{5.11}$$

where $\mathbf{V}(\omega)$ and $\mathbf{W}$ represents the matrix of cross-power-spectral density and initial weight matrix, respectively. The co-variance matrix for L number of frames is given by [91]:

$$\hat{\mathbf{V}}(\omega) = \frac{1}{L} \sum_{l=0}^{L-1} \mathbf{X}_l(\omega) \mathbf{X}_l^H(\omega), \tag{5.12}$$

where $\hat{\mathbf{V}}(\omega)$ is estimated co-variance matrix. The weights are optimized by minimizing the noise with the constraint of unity gain for the desired signal, i.e.,

$$\begin{aligned} &\underset{\mathbf{W}}{arg\ min} \quad \mathbf{W}^H(\omega)\hat{\mathbf{V}}(\omega)\mathbf{W}(\omega), \\ &subject\ to \quad \mathbf{W}^H(\omega)\mathbf{m} = 1, \end{aligned} \tag{5.13}$$

where $\mathbf{m}$ represents the steering vector, which is the most crucial matrix for direction estimation of the desired signal. It provides the directional information of microphone array. During this minimization, it affects the impulse response of the acoustic medium. Let d_i be the desired direction representation for the element i. Then, steering vector for i^{th} element (i.e., m_i) is given by:

$$m_i = e^{jwd_i}. \tag{5.14}$$

Constrained minimization in eq.(5.13) is performed by using Lagrange multipliers [96]. Hence, the optimum weight matrix for MVDR beamformer is given by:

$$\mathbf{W}_o(\omega) = \frac{\hat{\mathbf{V}}^{-1}(\omega)\,\mathbf{m}}{\mathbf{m}^H\,\hat{\mathbf{V}}^{-1}(\omega)\,\mathbf{m}}. \tag{5.15}$$

These optimum weights are utilized to obtain beamformed signal from the microphone array signal, i.e.,

$$\mathbf{D}(\omega) = \mathbf{W}_o^H(\omega)\mathbf{X}(\omega). \tag{5.16}$$

Furthermore, the output power ($\mathbf{p}_o$) of MVDR beamformer is given by:

$$\mathbf{p}_o(\omega) = \mathbf{W}_o^H(\omega)\hat{\mathbf{V}}(\omega)\mathbf{W}_o(\omega). \tag{5.17}$$

5.4 Reverberation Analysis Using Time-Domain Representation of Speech Signals

Fig. 5.2 shows the time-domain representation of genuine (Fig. 5.2(c)) and replay (Fig. 5.2(d)) signals. The Fig. 5.2(a) and Fig. 5.2(b) represents the zoomed version of the dotted squared region from Fig. 5.2(c) and Fig. 5.2(d), respectively. Furthermore, Fig. 5.2(e) and Fig. 5.2(f) corresponds to the zoomed version of the solid squared region from Fig. 5.2(c) and Fig. 5.2(d), respectively. Hence, from this zoomed figures, it can be observed that the replayed signal has additional impulses and distortions as compared to the genuine speech, which are due to the

added reverberation. This is in agreement with the other recent studies reported in [39, 97].

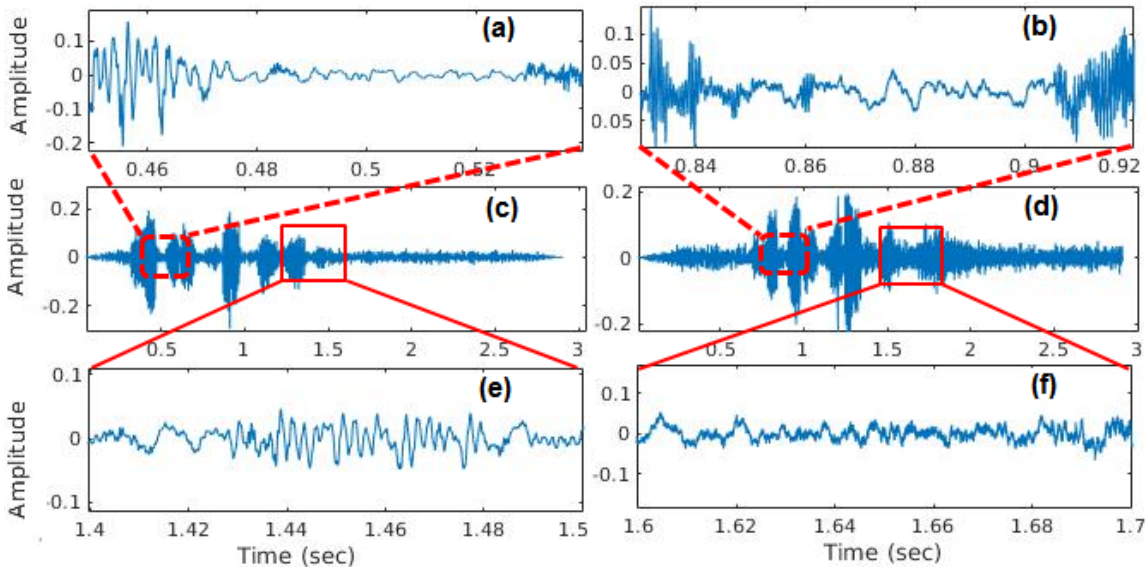


Figure 5.2: Time-domain representation of (c) genuine and (d) replayed speech signal from ReMASC dataset. Figure 5.2(a) and Figure 5.2(b) represents the zoomed version of the dotted squared region and Figure 5.2(e) and Figure 5.2(f) corresponds to the zoomed version of the solid squared region from Figure 5.2(c) and Figure 5.2(d), respectively.

5.5 Experimental Results

The performance of DAS *vs.* MVDR beamformer is evaluated using % EER. The SSD systems are developed for CQCC, LFCC, and TECC feature sets using GMM, CNN, and LCNN-based classifiers for all the three datasets, i.e., ReMASC and its DAS *vs.* MVDR beamformed versions. The % EER on development and evaluation sets are shown in Table 5.1 for all the three variants of datasets. It was observed that *only static* features performed better than all the other combinations. Hence, all the results reported in Table 5.1 are obtained using only static features. Furthermore, improved performance is obtained on the DAS beamformed dataset than that for the original ReMASC and MVDR beamformed version, for all the feature sets and classifiers considered in this study. This suggests that the DAS beamforming can be potentially utilized to improve the performance of the replay SSD system for VAs. In addition, the TECC feature set performs better than that of other feature sets for all the classifiers and all the dataset versions. This proves the capability of TECC to extract the reverberation characteristics in replay speech signal. In particular, relatively the best performance is observed for

TECC-GMM SSD system for DAS beamformed dataset. It should also be noted that, results of MVDR are worse even than unprocessed (i.e., not beamformed) ReMASC data indicating that MVDR is not all suitable beamforming for replay SSD task.

Furthermore, the performance of all the systems are also shown using Detection Error Trade-off (DET) curves in Figure 5.3. In particular, Figure 5.3(a) and Figure 5.3(b) shows the DET curves for development and evaluation set, respectively, for TECC-GMM system on all the three versions of datasets. It can be observed from Figure 5.3 that the DAS beamformed ReMASC consistently performing well as compared to ReMASC and its MVDR beamformed version for both development and evaluation sets.

Table 5.1: Results (in % EER) on ReMASC and its DAS vs. MVDR beamformed versions using various feature sets and classifiers. After [7].

Feature Set	Dataset	ReMASC		MVDR		DAS	
	Classifier	Dev.	Eval.	Dev.	Eval.	Dev.	Eval.
CQCC	GMM	19.94	22.56	36.74	30.73	**16.86**	**21.67**
	CNN	15.36	25.33	30.84	29.95	**12.12**	**22.38**
	LCNN	17.85	27.64	34.30	32.80	**15.25**	**24.78**
LFCC	GMM	22.39	23.38	35.53	30.47	**20.06**	**21.66**
	CNN	15.04	25.27	28.67	28.12	**12.13**	**20.13**
	LCNN	15.69	24.96	35.70	32.65	**16.66**	**22.96**
TECC	GMM	20.42	17.75	36.13	26.61	**16.52**	**14.94**
	CNN	15.80	23.99	31.16	28.82	**13.31**	**21.74**
	LCNN	15.90	23.86	36.03	31.56	**14.71**	**22.56**

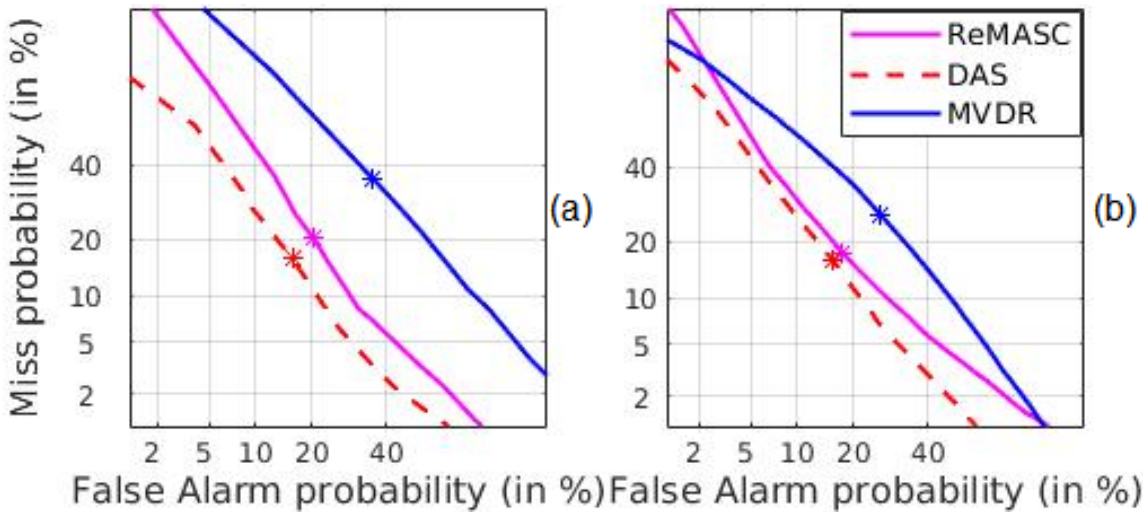


Figure 5.3: DET curves for ReMASC and its beamformed versions using TECC with GMM: (a) development set, and (b) evaluation set. After [7]

5.5.1 Analysis of Latency Period

In this study, we have analyzed the trade-off between % EER and latency period (as shown in Figure 5.4), using TECC-GMM SSD system for development and evaluation sets of ReMASC and its beamformed versions (i.e., DAS and MVDR). The latency period of the trained model is estimated by computing the % EER *w.r.t.* varying durations of test speech segment in a test utterance. For latency period analysis, we chose the duration of the utterances varying from 20 ms to 2000 ms with an interval of 200 ms. It can be observed from the Figure 5.4 that even for a short latency period, DAS is performing better than the other two versions of datasets and hence, it shows the significance of DAS beamformer for practical SSD system deployment for VAs.

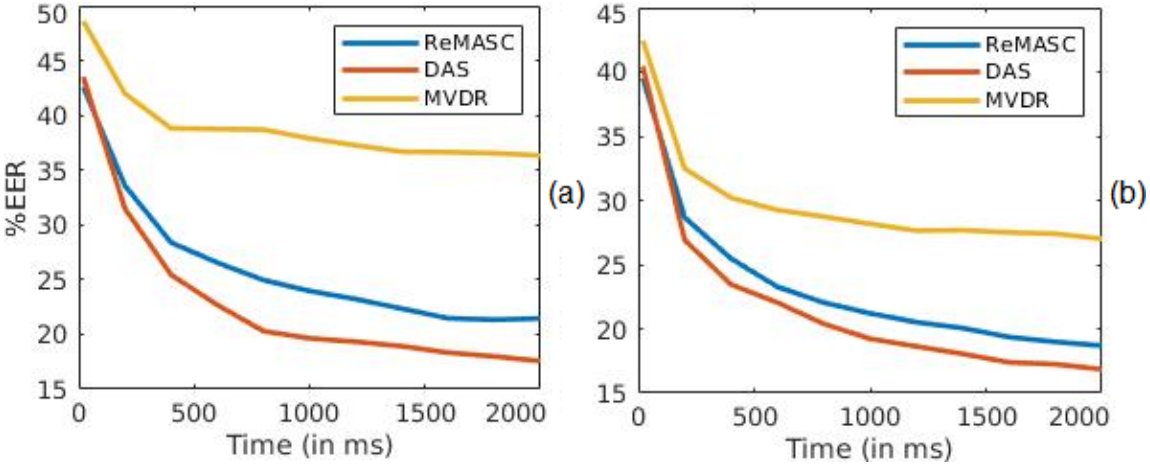


Figure 5.4: Latency period analysis for TECC-GMM SSD system on ReMASC and its DAS and MVDR beamformed versions. After [7].

5.6 Chapter Summary

This chapter presented the significance of DAS beamformer over MVDR for replay SSD task on VAs. This crucial observation found in this work is contradictory w.r.t. suitability of state-of-the-art MVDR beamformer for Distant Speech Recognition (DSR), indicating straightforward generalization of beamforming method from DSR to replay SSD in VAs is not recommended even through DSR is very much integral part of VAs. In addition, due to linear phase characteristics of DAS beamformer, the acoustical characteristics of reverberation in replay spoof are presented and hence, TECC is employed to capture these reverberation characteristics. Performance comparison with the existing CQCC and LFCC feature

sets indicates better performance offered by TECC. Finally, analysis of latency indicates potential of DAS beamformer w.r.t. TECC-GMM for practical SSD system deployment. Our future work will be directed to extend this work on the other beamforming techniques, with the aim of capturing reverberation along with the least possible latency.

CHAPTER 6
Wavelet-Based Features for VLD

6.1 Introduction

In this chapter, the Morlet wavelet-based features for VLD via pop noise detection is proposed. With respect to Heisenberg's uncertainty principle in signal processing framework, wavelet-based approach offers improved resolution in time and frequency as compared to the STFT-based method. Furthermore, Morlet wavelets are known to capture perceptual cues effectively (both in visual and hearing domains). To that effect, the use of Morlet wavelet to capture discriminating cues based on pop noise for genuine *vs.* replay spoof classification is being proposed for the first time in this thesis. Experiments are presented for two CWT-based features, namely, Handcrafted Morlet Wavelet and Low Frequency Morlet Scalogram-based Features on POp noise COrpus (POCO) for VLD. [1]

6.2 Continuous Wavelet Transform (CWT)

The effect of *human breath* on a microphone results in a sudden high energy (i.e., pop noise as an event in speech) in low frequency regions. To locate pop noise, time-frequency representations, such as spectrograms, have been used in the recent past [21, 53]. However, to get better detection of pop noise, we have used CWT in this work. The key idea behind employing wavelet for pop noise detection is to exploit the capability of a wavelet (which is a wave for a short duration) to capture *transients* in a speech wave, i.e., occurrence of pop noise. A mother wavelet $\psi(t) \in L^2(R)$ (i.e., Hilbert space of finite energy signals) is a wave of short duration that has zero average. It is defined as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\Big(\frac{t-b}{a}\Big), \qquad a \in R^+, b \in R, \tag{6.1}$$

[1] This work is jointly done with PhD. scholar at DA-IICT, Ms. Priyanka Gupta.

where b is called the translation (position), and a is called the dilation (scale) coefficient. There are various types of wavelets. The most famous wavelet is the Morlet wavelet, which is a modulated Gaussian, and it is defined as [98]:

$$\psi(t) = e^{j\omega_0 t} e^{-t^2/2}, \tag{6.2}$$

where ω_0 is taken as 5 Hz for a standard Morlet wavelet. The Morlet wavelet is obtained from a Gaussian window multiplied by a sinusoidal wave [99]. The CWT of signal $f(t)$ is

$$\begin{aligned} Wf(a,b) &= < f(t), \psi_{a,b}(t) >, \\ Wf(a,b) &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi^* \Big(\frac{t-b}{a} \Big) dt, \end{aligned} \tag{6.3}$$

where $< \cdot, \cdot >$ indicates inner product operation to compute wavelet coefficients, and $*$ denotes complex conjugate. We have considered Morlet wavelet in this work because it is closely related to human perception (for both hearing and vision) [100]. Moreover, CWT is related to constant-Q filtering- a short-time analysis performed by the peripheral auditory system. In particular, as per the original investigations by Flanagan [101], the wavelet function, for the mechanical spectral analysis performed by the basilar membrane in the human ear is given by $\psi(t) = (t\omega)^2 e^{-t\omega/2}$ [101]. Furthermore, Morlet wavelet is the most widely used wavelet for CWT and, in fact, the first wavelet of its kind in formal historical developments of wavelets in the geophysics literature for detection of transients and improving joint time-frequency resolution of seismic signals [102].

6.3 Proposed Approaches

The feature extraction for Spoofed Speech Detection (SSD) task is based on the hypothesis that both genuine and spoof utterances possess differences w.r.t. presence and absence of pop noise energy levels, respectively. Fig. 6.1 shows the scalograms of the word 'laugh'. A distinct signature of pop noise can be seen in Panel I. However, the pop noise signature is not so distinct for the case, when a pop filter was used as shown in Panel II of Figure 6.1.

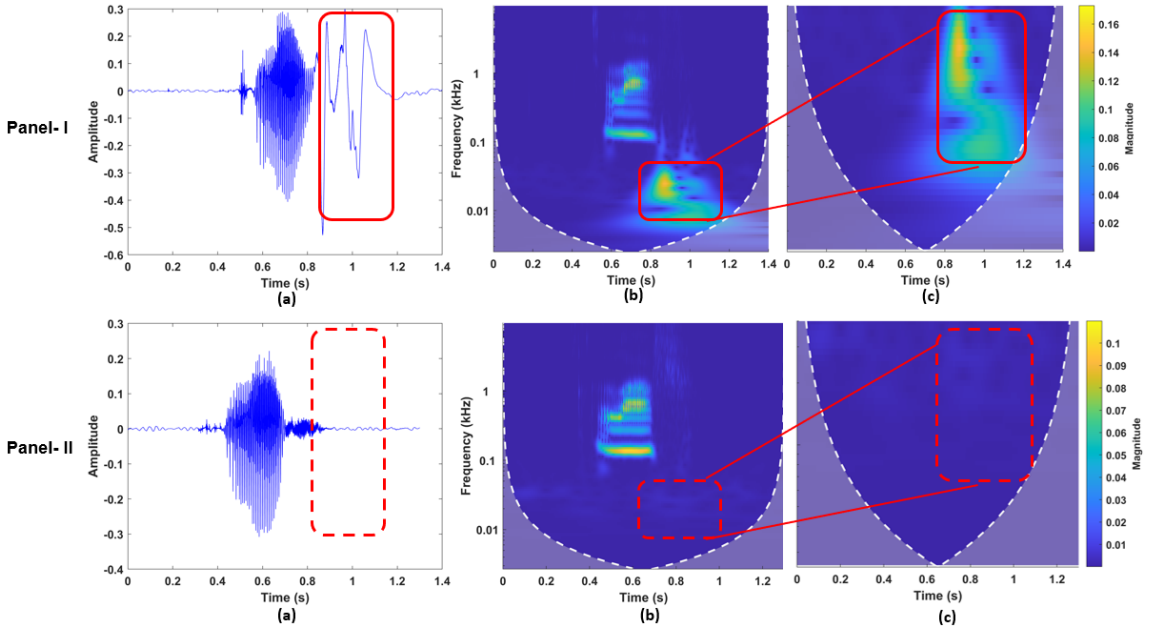


Figure 6.1: Panel I represent the case of presence of pop noise (genuine speech) indicated by box. Panel II represents the case of reduced pop noise (spoofed speech) due to the use of pop filter, (a) time-domain signal for the word *'laugh'*, (b) corresponding scalogram, and (c) corresponding low frequency ($0 - 40$ Hz) scalogram. Solid boxes in Panel I indicate the presence of pop noise, while corresponding dotted boxes in Panel II indicates that the pop noise has been eliminated due to pop filter.

6.3.1 Handcrafted Morlet Wavelet-Based Features

CWT coefficients are extracted from the speech data of POCO corpus by taking Morlet as the mother wavelet. CWT coefficients are found for frequencies $\leq$ 40 Hz, as shown in Algorithm 2. Furthermore, to keep the dimension (D) of the feature vector as 45 and also to extract the prominent energy of pop noise, the energies are arranged in descending order, and the highest *45*-D values are taken for extracting the *45*-dimensional feature vector.

6.3.2 Low Frequency Morlet Scalogram-Based Features

Scalogram is a visual time-frequency representation of CWT coefficients. In particular, it can be interpreted as a time-frequency energy density, $|Wf(a,b)|^2$ [99]. The time-frequency resolution of the wavelet transform depends on the frequency of the signal. At high frequencies, the wavelet reaches a high time resolution but a low frequency resolution. At low frequencies, high frequency resolution and low time resolution can be obtained. Since pop noise most likely occurs at frequency regions $\leq$ 40 Hz, scalograms are very well suited to extract energies at

Algorithm 2: Proposed Handcrafted Morlet Wavelet-based Feature Extraction for VLD. After [8].

Speech signal $f(t)$ Feature w_name='amor' `// Taking Morlet wavelet`
[cwt_coeffs, F] ← cwt(f(t), w_name)
`/* Finding CWT coefficients for low frequencies */`
Low_F ← find ($0 < F \leq 40$ Hz)
  Low_coeffs ← cwt_coeffs (Low_F)
Pop_energy = abs (Low_coeffs)^2
`/* Converting pop energy to a 45-D feature vector */`
dim ← 45
M=mean (Pop_energy)
SD=standard_deviation (Pop_energy)
k ← length (Low_coeffs)
while $k > 0$ **do**
  $i = 1$
  $Norm\_Pop(i) = \frac{Pop\_energy(i) - M}{SD}$
  $k--, i++$
[sorted, index] ← sort (Norm_Pop, descending)
Feature ← Pop_energy (index(1:dim))

low frequencies because of the higher frequency resolution of scalogram at lower frequencies.

For our experiments, the lowest frequency bin is set at 1.9826 Hz. The scale factor between 2 consecutive bins is 1.0718. Therefore, the k^{th} bin index corresponding to 40 Hz is calculated as:

$$40 = (1.0718)^k * 1.9826. \tag{6.4}$$

Therefore, frequency region approximately below 40 Hz is found to be corresponding to the nearest integer $k = 44$ frequency bins. Taking bin index below $k = 44$, we get frequencies exactly below 41.9025 Hz. This is the region where the pop noise is located. To that effect, scalogram images are extracted only corresponding to 44 wavelet coefficients. Each scalogram image is of the size 512×512. These scalogram-based features are then fed as an input to the CNN.

6.4 Baseline Approaches

6.4.1 Low Frequency Spectrogram-Based Features

Low frequency spectrogram-based features for VLD were extracted from the STFT in [21]. The same algorithm was used on POCO dataset in [5]. In this work, energies only in the low frequency (in particular, < 40 Hz) regions were extracted by selecting frequency bins corresponding to 0 to 40 Hz. Next, the average S_{eng} of the spectral energy densities of the STFT-based spectrogram was calculated by averaging across the bins for every k^{th} frame. For the framewise spectral energies obtained in S_{eng}, mean and standard deviation were calculated to obtain normalized values. The frames with the 10 highest energies were selected to get meaningful spectrogram-based features for pop noise detection. The classifier used was Support Vector Machines (SVM).

6.4.2 CQT-Based Features

An improvement to the baseline was introduced in [56], using CQT-based features. As compared to the STFT that has constant frequency resolution, CQT has geometrically distributed frequency bins due to constant-Q ratio of center frequency to resolution. The number of bins per octave is taken to be 96 and the number of samples taken in the first octave is 2. Furthermore, f_{min} is set to 0.48 Hz and f_{max} is set to 11050 Hz. For classification, the study reported in [56] used SVM-based classifier.

6.4.3 Mel Spectrogram-Based Features

Apart from our proposed CWT-based approach in this work, we also include the use of Mel spectrogram (to our knowledge, this is not utilized for VLD task in the literature) for the purpose of comparing our experimental results. We estimated pop noise energies using the STFT-based approach on Mel Spectrogram only on frequencies < 40 Hz. Therefore, we estimated the Mel spectrogram with 16 number of bands and 5400 as the FFT length for better frequency resolution. Classification was done using a CNN-based classifier described in sub-Section 3.4.1.

6.5 Experimental Results

This Section describes the experimental results for the proposed CWT-based features for VLD task. Further analysis is done to investigate the effect of various phonemes on the accuracy, by finding wordwise accuracies.

6.5.1 Proposed Handcrafted Morlet-Based Features

For the case of *45*-D wavelet-based features (shown as system (F)), we achieved an overall accuracy of 80 %. Fig. 6.2 shows wordwise accuracy over 44 words in the dataset. We observed that the word 'pay' has the highest accuracy of 91.02 %, because the word 'pay' has a strong plosive sound of /p/. Furthermore, we achieved an average accuracy of 79.35 % and 79.27 % on words with prominent performance on plosive and fricatives, respectively, as shown in Table 6.1.

6.5.2 Proposed Morlet Scalogram-Based Features

The Morlet scalogram features (shown as system (G)) performed significantly well as compared to the traditional STFT-based baseline system. We observed overall accuracy of 86.23% on Morlet scalogram-based features. We observed that the word 'tourist' has the highest accuracy of 97.43%, because the word 'tourist' has 2 strong plosive sounds of /t/. Given the effect of pop noise depends on the uttered word, we achieved an average accuracy of 89.07% and 87.61% on words with prominent plosive and fricatives, respectively.

Table 6.1: Average accuracy (in %) of different phoneme types. After [8].

Phoneme Type	**(A) Spectrogram (SVM) [5]**	**(B) CQT (SVM) [56]**	**(C) Spectrogram (CNN) [55]**	**(D) Mel-spectrogram (CNN)**	**(E) Handcrafted Bump Wavelet-based (CNN) [57]**	**(F) Handcrafted Morlet Wavelet-based (CNN) (Proposed)**	**(G) Handcrafted Morlet Scalogram (CNN) (Proposed)**
Freq. Range	**0-40 Hz**	**0-11025 Hz**	**0-11025 Hz**	**0-40 Hz**	**0-40 Hz**	**0-40 Hz**	**0-40 Hz**
Plosive	60.46	63.61	71.72	74.13	81.58	79.35	**89.07**
Fricatives	67.66	73.78	75.55	77.45	80.77	79.27	**87.61**
Whisper	68.44	73.29	76.83	74.99	81.09	79.48	**86.21**
Nasal	54.26	57.78	59.33	70.51	76.50	71.36	**80.77**
Liquids	69.78	57.16	56	69.23	69.87	65.38	**79.49**
Affricates	58.26	68.92	71.83	72.51	78.53	74.35	**85.26**

6.5.3 Discussion

It can be observed in Table 6.1 that our proposed Morlet scalogram-based approach outperforms every other methods for *all* the phoneme types. Further-

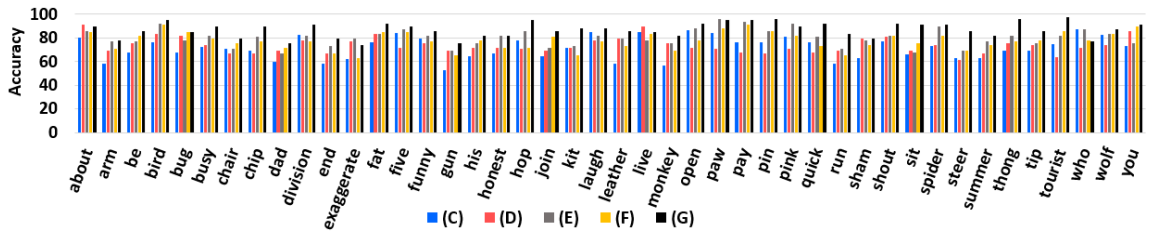


Figure 6.2: Word wise accuracies (in %) with CNN classifier for (C): Full-frequency spectrogram, (D): Low-frequency Mel-spectrogram, (E): Handcrafted Bump wavelet-based features, (F): Handcrafted Morlet wavelet-based features, and (G): Handcrafted Morlet scalogram. After [8].

more, we also observe that all the methods perform relatively better for plosive and fricative sounds. Fricative sounds (such as, /f/ sound in the word 'lau**gh**') are produced due to turbulent airflow, which results in bursts of energy at low frequencies for a short-time period, characterizing the presence of pop noise. Furthermore, plosive sounds (such as, /p / sound in '**p**ay') are caused by a sudden release of a burst of air from the lips, resulting in pop noise [75]. On the contrary, energy distribution in nasal sounds is due to partial air released from the nostrils and the mouth [75]. Since the released air is coming from two sources, it barely results in energy at low frequency regions. To that effect, the accuracy score of all the algorithms are relatively lower for the nasal sounds.

6.6 Chapter Summary

In this chapter, the CWT is used to effectively improved resolution in time and frequency for VLD-based on pop noise. VLD enables to discriminate a *live* voice from the other *non-live* voice signals, such as replayed, voice converted, and synthetically generated signals. To that effect, two handcrafted features were proposed in this study: Morlet wavelet-based features, and Morlet scalogram-based features. A significant improvement in accuracy is observed with both the features as compared to the existing systems. Further analysis shows the effect of phoneme type on the accuracy. However, the proposed approach comes with a trade-off between high performance and computational complexity. Further, similar wavelet-based methodologies can be tested for various configurations of spoof signals, as future work. Furthermore, the combined effect of microphone variability on ASV and pop noise-based VLD task can also be investigated. In the next chapter, the entire thesis is summarized, along with limitations and potential research directions.

CHAPTER 7
Summary and Conclusions

This thesis investigated the significance of quadrature phase for subband signals of cochlear filterbank in the framework of CFCCIF-QESA feature extraction. Significance of quadrature phase is motivated from the analysis of mutual information (MI) for AM-FM signal and in the original definition of Fourier transform. In order to incorporate quadratic phase, an analytic signal was generated using Hilbert transform. The key idea behind this was to exploit the advantages of relative phase information and excellent temporal resolution of ESA to estimate IF for each subband signal to derive the CFCCIF-QESA feature set. The development of CFCCIF-QESA was motivated by the success of CFCC for mismatched training and testing conditions and recent success of CFCCIF as winner system for ASVSpoof 2015 challenge campaign and its recent follow-up work on using CFCCIF-ESA for replay SSD tasks on ASVSpoof 2017 $v2.0$ dataset. The performance of the proposed feature set was evaluated on ASVSpoof 2017 $v2.0$ and BTAS 2016 datasets (both development and evaluation sets). The key motivation for these intensive experiments on various datasets was to explore the generalizability of the proposed feature set to other datasets.

The another study presented significance of DAS beamformer over MVDR for replay SSD task on VAs. This crucial observation found in this work is contradictory w.r.t. suitability of state-of-the-art MVDR beamformer for Distant Speech Recognition (DSR), indicating straightforward generalization of beamforming method from DSR to replay SSD in VAs is not recommended even through DSR is very much integral part of VAs. In addition, due to linear phase characteristics of DAS beamformer, the acoustical characteristics of reverberation in replay spoof are presented and hence, TECC is employed to capture these reverberation characteristics. Performance comparison with existing CQCC and LFCC indicates better performance offered by TECC. Finally, analysis of latency indicates potential of DAS beamformer w.r.t. TECC-GMM for practical SSD system deployment.

In the work done for VLD, we used CWT to effectively improved resolution

in time and frequency. VLD enables to discriminate a *live* voice from the other *non-live* voice signals, such as replayed, voice converted, and synthetically generated signals. To that effect, two handcrafted features were proposed in this study: Morlet wavelet-based features, and Morlet scalogram-based features. A significant improvement in accuracy is observed with both the features as compared to the existing systems. Further analysis shows the effect of phoneme type on the accuracy.

7.1 Limitations of the Thesis Work

- The proposed CFCCIF-QESA features does not yield improved performance for neural network-based classifiers. Further parameter tuning can be done to improve the performance for CNN and LCNN-based classifiers.
- proposed CFCCIF-QESA features do not yield improved performance than CQCC baseline for ASVspoof 2019 PA dataset, because it contains simulated replay utterances, unlike ASVspoof 2017 dataset which contains replay utterances under realistic scenarios.
- The wavelet-based features gave relatively high accuracy as compared to the baseline approach for VLD. However, the proposed approach comes with a trade-off between high performance and computational complexity.

7.2 Future Research Directions

- The upcoming research efforts will be directed towards investigating the significance of the proposed CFCCIF-QESA feature set on the other spoofing attacks, such as VC and SS on ASVSpoof 2015, ASVSpoof 2019 challenge dataset, and on the recently released DeepFake speech data of ASVSpoof 2021 challenge.
- Apart from CNN and LCNN used in this work, one can investigate the other deep learning-based classifiers, such as ResNet and LSTM.
- The cross-database evaluation can be done to verify the generalizability of CFCCIF-QESA feature set.
- For VLD systems, similar wavelet-based methodologies as proposed in this work can be tested for various configurations of spoof signals. Furthermore,

the combined effect of microphone variability on ASV and pop noise-based VLD task can also be investigated.

- The future research direction for replay SSD on VAs is to extend presented work on the other beamforming techniques, with the aim of capturing reverberation along with the least possible latency.

List of Publications

Journal

1. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Replay Spoof Detection Using Energy Separation Based Instantaneous Frequency Estimation From Quadrature and In-Phase Components," **accepted** in Computer, Speech & Language, Elsevier, (2022).

2. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Synthetic and Voice Converted Spoof Detection Using Energy Separation Based Instantaneous Frequency Estimation From Quadrature and In-Phase Components," **article under preparation** in Computer, Speech & Language, Elsevier, (2022).

Conferences

1. Anand Therattil, Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Teager Energy Based-Detection of One-point and Two-point Replay Attacks: Towards Cross-Database Generalization," **accepted** in Speaker Odyssey-The Speaker Recognition Workshop, Beijing, China, June 28 - July 01, 2022.

2. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Energy Separation Based Instantaneous Frequency Estimation from Quadrature and In-Phase Components for Replay Spoof Detection," **accepted** in 30^{th} European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, August 29-September 2, 2022.

3. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Morlet Wavelet-Based Voice Liveness Detection using Convolutional Neural Network," **accepted** in 30^{th} European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, August 29-September 2, 2022.

4. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Morse Wavelet Features for Pop Noise Detection," **accepted** in International Conference on Signal Processing and Communications (SPCOM), Bengaluru, India, 11-15 July 2022.

5. **Piyushkumar K. Chodingala**, Shreya S. Chaturvedi, Ankur T. Patil, Hemant A. Patil, "Robustness of DAS Beamformer Over MVDR for Replay Attack Detection On Voice Assistants," **accepted** in International Conference on Signal Processing and Communications (SPCOM), Bengaluru, India, 11-15 July 2022.

6. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, "Relevance of Quadrature Phase For Replay Detection in Voice Assistants (VAs)," **rejected** in International Conference on Signal Processing and Communications (SPCOM), Bengaluru, India, 11-15 July 2022.

7. **Piyushkumar K. Chodingala**, Shreya S. Chaturvedi, Ankur T. Patil, Hemant A. Patil, "DAS *vs.* MVDR: Which Beamformer is Suitable For Replay Attack Detection On Voice Assistants?," **rejected** in INTERSPEECH, Incheon, Korea, September 18 to 22, 2022.

8. Priyanka Gupta, **Piyushkumar K. Chodingala**, Hemant A. Patil, Ankur T. Patil, "Significance of Quadrature and In-Phase Components for Synthetic Spoofed Speech Detection," **rejected** in INTERSPEECH, Incheon, Korea, September 18 to 22, 2022.

9. **Piyushkumar K. Chodingala**, Ankur T. Patil, Hemant A. Patil, Shreya S. Chaturvedi "Significance of DAS *vs.* MVDR Beamformer for Replay Spoof Detection On Voice Assistants," **rejected** to 30^{th} European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, August 29-September 2, 2022.

10. Ankur T. Patil, Hemant A. Patil, Priyanka Gupta, **Piyushkumar K. Chodingala**, "Corpora for Voice Anti-Spoofing Research and Development," **rejected** in Speaker Odyssey-The Speaker Recognition Workshop, Beijing, China, June 28 - July 01, 2022.

References

[1] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, Les Sables dĂŹOlonne, France, 26 - 29 June, 2018, pp. 296–303.

[2] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Biometrics: Theory, Applications, and Systems (BTAS)*, Arlington, USA, 2015, pp. 1–6.

[3] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. S. Mello, R. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition," in *Biometrics: Theory, Applications, and Systems (BTAS)*, Niagara Falls, Buffalo, USA, Sept. 2016, pp. 1–6.

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

[5] K. Akimoto, S. P. Liew, S. Mishima, R. Mizushima, and K. A. Lee, "POCO: a voice spoofing and liveness detection corpus based on pop noise," *INTERSPEECH*, pp. 1081–1085, Shanghai, China, 25-29 Oct., 2020.

[6] P. Gupta, P. K. Chodingala, and H. A. Patil, "Energy separation based instantaneous frequency estimation from quadrature and in-phase components for replay spoof detection," in *revised and submitted major revision in Computer Speech & Language*, may 14 2022.

[7] P. K. Chodingala, S. S. Chaturvedi, A. T. Patil, and H. A. Patil, "DAS *vs.* MVDR: Which Beamformer is Suitable For Replay Attack Detection On Voice Assistants?" in *submitted in INTERSPEECH, Incheon, Korea*, September 18 to 22, 2022.

[8] P. Gupta, P. Chodingala, and H. A. Patil, "Morlet wavelet-based voice liveness detection using convolutional neural network," in *submitted in European Signal Processing Conference (EUSIPCO), Belgrade, Serbia*, August 29-September 2, 2022.

[9] A. Meyer, D. Döbler, J. Hambrecht, and M. Matern, "Acoustic mapping on three-dimensional models," in *Proceedings of the* 12^{th} *International Conference on Computer Systems and Technologies*, Vienna, Austria, 2011, pp. 216–220.

[10] "HSBC reports high trust levels in biometric tech as twins spoof its voice ID system," *Biometric Technology Today*, vol. 2017, no. 6, 2017.

[11] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[12] Y. Stylianou, "Voice transformation: A survey," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 19-24 April, 2009, pp. 3585–3588.

[13] 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," pp. 1–6, *INTERSPEECH 2017*, Stockholm, Sweden, 20-24 August 2017.

[14] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. INTERSPEECH, Lyon, France*, 25-28 August 2013, pp. 925–929.

[15] *Spoofing and Countermeasures for Automatic Speaker Verification, Special Sessions in INTERSPEECH-2013*, Last Accessed July 25, 2020. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2013/i13_specials.pdf

[16] 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," in *INTERSPEECH*, Dresden, Germany, 6-10 September 2015, pp. 2037–2041.

[17] A. et al., "ASVSpoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.

[18] 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 Trans. on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[19] V. Kepuska and G. Bohouta, "Next generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa, and Google home)," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), University of Nevada, United States*, 8-10 Jan. 2018, pp. 99–103.

[20] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic replay attack corpus for voice controlled systems," *INTERSPEECH, Graz, Austria*, pp. 2355–2359, 15-19 September 2019.

[21] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2047–2051.

[22] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu, "Secure your voice: An oral airflow-based continuous liveness detection for voice assistants," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, NY, USA*, vol. 3, no. 4, pp. 1–28, 2019.

[23] S. Mochizuki, S. Shiota, and H. Kiya, "Voice livness detection based on pop-noise detector with phoneme information for speaker verification," *The Journal of the Acoustical Society of America (JASA)*, vol. 140, no. 4, pp. 3060–3060, 2016.

[24] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 44–56, 2017.

[25] Z. Wu, A. Larcher, K. A. Lee, E. S. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: The effect of text constraints," in *Proc. INTERSPEECH, Lyon, France*, 25-28 August 2013, pp. 950–954.

[26] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVSpoof 2015

challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China*, 20-25 March 2016, pp. 5475–5479.

[27] M. Wester, Z. Wu, and J. Yamagishi, "Human *vs.* machine spoofing detection on wideband and narrowband data," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2047–2051.

[28] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *INTERSPEECH*, Dresden, Germany, 6-10 September 2015, pp. 2092–2096.

[29] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2082–2086.

[30] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high-dimensional magnitude and phase features: The NTU approach for ASVSpoof 2015 challenge," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2052–2056.

[31] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *IEEE Journal of Selected Topics in Sig. Process.*, vol. 11, no. 4, pp. 618–631, 2016.

[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[33] 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, 2017.

[34] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the ASVSpoof 2017 challenge," in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 7–11.

[35] M. Witkowski and et. al., "Audio replay attack detection using high frequency features,," in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 27–31.

[36] X. Wang, Y. Xiao, and X. Zhu, "Feature Selection Based on CQCCs for Automatic Speaker Verification Spoofing," in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 32–36.

[37] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "ASVSpoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2021. [Online]. Available: https://www.asvspoof.org/workshop

[38] L. Huang and J. Zhao, "Audio replay spoofing attack detection using deep learning feature and long-short-term memory recurrent neural network," in *AIIPCC 2021; The Second International Conference on Artificial Intelligence, Information Processing and Cloud Computing*. VDE, 2021, pp. 1–5.

[39] Madhu R. Kamble and Hemant A. Patil, "Detection of replay spoof speech using teager energy feature cues," *Computer Speech & Language*, vol. 65, p. 101140, 2021.

[40] A. T. Patil, R. Acharya, H. A. Patil, and R. C. Guido, "Improving the potential of enhanced Teager energy cepstral coefficients (ETECC) for replay attack detection," *Computer Speech & Language*, vol. 72, p. 101281, 2022.

[41] 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*, 2017, pp. 17–21.

[42] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep siamese architecture based replay detection for secure voice biometric." in *INTERSPEECH, Hyderabad, India*, 2-6 September 2018, pp. 671–675.

[43] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using dnn for channel discrimination." in *INTERSPEECH, Stockholm, Sweden*, August 20-24, 2017, pp. 97–101.

[44] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention." in *INTERSPEECH, Hyderabad, India*, 2-6 September 2018, pp. 681–685.

[45] Tanvina B. Patel and Hemant A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *INTERSPEECH, Dresden, Germany*, 6-10 September 2015, pp. 2062–2066.

[46] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Tran. on Sig. Process.*, vol. 41, no. 10, pp. 3024–3051, 1993.

[47] A. T. Patil, R. Acharya, P. K. A. Sai, and H. A. Patil, "Energy separation-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection," in *INTERSPEECH, Graz, Austria*, 15-19 September 2019, pp. 2898–2902.

[48] H. M. Teager, "Some observations on oral airflow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[49] James F Kaiser, "On a simple algorithm to calculate the energy'of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, USA*, 3-6 April 1990, pp. 381–384.

[50] 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)*, 2021, pp. 6364–6368.

[51] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." in *INTERSPEECH, Stockholm, Sweden*, August 20-24, 2017, pp. 82–86.

[52] L. Huang and C.-M. Pun, "Audio replay spoof attack detection using segment-based hybrid feature and densenet-lstm network," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019, pp. 2567–2571.

[53] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE Conference on Computer Communications*, 29 April - 2 May 2019, Paris, France, pp. 2062–2070.

[54] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection using phoneme-based pop-noise detector for speaker verification," in *Odyssey 2018 The Speaker and Language Recognition Workshop. ISCA*, 2018, pp. 233–239.

[55] Siddhant Gupta, Kuldeep Khoria, Ankur T. Patil and Hemant A. Patil, "Deep Convolutional Neural Network for Voice Liveness Detection," in *Speech and Computer International Conference (SPECOM)*. Springer, 2021.

[56] K. Khoria, A. T. Patil, and H. A. Patil, "Significance of constant-q transform for voice liveness detection," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 126–130.

[57] P. Gupta, Siddhant Gupta, and Hemant A. Patil, "Voice Liveness Detection using Bump Wavelet with CNN," in *International Conference on Pattern Recognition and Machine Intelligence (LNCS)*. Springer, 15-18 December 2021.

[58] S. Singh, K. Khoria, and H. A. Patil, "Modified group delay function using different spectral smoothing techniques for voice liveness detection," in *IEEE International Conference on Signal Processing and Communications (SPECOM), Petersburg, Russia*. Springer, 2021, pp. 649–659.

[59] 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, no. 4, pp. 588–604, 2017.

[60] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The REDDOTS data collection for speaker recognition," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2996–3000.

[61] A. T. Patil, H. Kotta, R. Acharya, and H. A. Patil, "Spectral root features for replay spoof detection in voice assistants," in *SPECOM, Petersburg, Russia*. Springer, 2021, pp. 504–515.

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

[63] P. Gupta, S. Gupta, and H. A. Patil, "Voice liveness detection using bump wavelet with CNN," in *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, December 15 - 18, 2021.

[64] K. Khoria, A. T. Patil, and H. A. Patil, "Significance of constant-Q transform for voice liveness detection," in 29^{th} *European Signal Processing Conference (EUSIPCO), Dublin, Ireland*, 2021, pp. 126–130.

[65] D. A. Reynolds, "Gaussian mixture models." *Encyclopedia of Biometrics*, vol. 741, pp. 659–663, 2009.

[66] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[67] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning (ICML), Haifa, Israel*, 21-24 June 2010.

[68] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of The Trade*. Springer, 2012, pp. 421–436.

[69] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016, vol. 1, no. 2.

[70] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[71] R. D. Nindrea, T. Aryandono, L. Lazuardi, and I. Dwiprahasto, "Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis," *Asian-Pacific Journal of Cancer Prevention: APJCP*, vol. 19, no. 7, p. 1747, 2018.

[72] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. *I.* Fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, 1992.

[73] J. F. Kaiser, "On teager's energy algorithm and its generalization to continuous signals," in *Proc.* 4^{th} *IEEE Digital Signal Processing Workshop*. Mohonk (New Palts), NY, 1990, pp. 338–375.

[74] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators." in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toronto, Ontario, Canada*, 14-17 April 1991, pp. 421–424.

[75] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 2^{nd} Edition, Pearson Education India, 2004.

[76] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[77] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[78] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[79] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *INTERSPEECH, Portland, USA*, 09 September 2012.

[80] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2011.

[81] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. on Info. Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[82] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal (BSTJ)*, vol. 27, no. 3, pp. 379–423, 1948.

[83] R. Hamila, J. Astola, F. A. Cheikh, M. Gabbouj, and M. Renfors, "Teager Energy and the Ambiguity Function," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 260–262, 1999.

[84] P. Maragos and A. C. Bovik, "Image demodulation using multidimensional energy separation," *Journal of the Optical Society of America A (JOSA)*, vol. 12, no. 9, pp. 1867–1876, 1995.

[85] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[86] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.

[87] A. V. Oppenheim, R. W. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 437–466, 1968.

[88] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 10-12 September, 2014, pp. 1–6.

[89] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Simon & Schuster, Inc., 1992.

[90] M. Karaman, P.-C. Li, and M. O'Donnell, "Synthetic aperture imaging for small scale systems," *IEEE Transactions on Ultrasonic, Ferroelectrics, and Frequency Control*, vol. 42, no. 3, pp. 429–442, 1995.

[91] L. G. Bezanson, *The Subarray MVDR Beamformer: A Space-Time Adaptive Processor Applied to Active Sonar*. University of California, San Diego, 2013.

[92] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley & Sons, 2009.

[93] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT Press Cambridge, MA, 1949, vol. 113, no. 21.

[94] E. A. Habets, J. Benesty, S. Gannot, and I. Cohen, "The MVDR beamformer for speech enhancement," in *Speech Processing in Modern Communication,*. Springer, 2010, pp. 225–254.

[95] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[96] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.

[97] 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), Brighton, UK*, 12-17 May 2019, pp. 2607–2611.

[98] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.

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

[100] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions On Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[101] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer Science & Business Media, 2013, vol. 3.

[102] I. Daubechies, "Where do wavelets come from? A personal point of view," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 510–513, 1996.

CHAPTER A
MATLAB Pseudo Code

MATLAB code of main CFCCIF-QESA feature set

```
% clear all;clc;close all;
% main QESA feature set
function feat = feat_cfccif_complex_esa(x,fs)

    P=1;    L=0;     H=fs/2;    Q=80;     Nc=12;
    fc=linspace(L,H,Q+2);       % make a Linear scale
    fc=fc(2:end-1);             % the central frequency of the ...
       filters
    fL=fc(1);                   % lowest cochlear filter's fc

    a=fL./(fc);                 % scaling parameter => to shift ...
       the center freq.
    %t=0:1/fs:0.0375;               % The time variable upto ...
       half second
    t=0:1/fs:0.05;
    d=floor(20*fs/1000);        % d=Window length(in ...
       samples/ms)floor(d/2);
    L=floor(8*fs/1000);               % shift of window (in ...
       samples/ms)
                    % No. of
    % Parameters for the shape of filter and Theta

    A=3;                           % Alpha value
    B=0.016;
    thetamin=Ttheta(Q,A,B,fL,t);
    %% Design the cochlear Filters
    % Start proceesing for each filter bank output
    for i=1:Q

        c1 = 1./sqrt(a(i));
        c2 = (t/a(i)).^A.*exp(-2*pi*fL*B*(t/a(i)));
```

```
        c3 = cos(2*pi*fL*(t/a(i))+thetamin);
        Si = c1.*c2.*c3;

        Tab = conv(x,Si);
        CTab=hilbert(Tab);
        %% The function of implementation of 1D motion of BM ...
            (unidirectional)
        hab = (Tab).^2;
        %% Hair cell output of each band==> i.e., representation ...
            of nerve spike density


        % Ti=3.5*(1./fc)*fs; d=max(3.5*ti,d)

        j=1;
        ss=zeros();
    inst_freq=zeros();

    for l=1:L:(length(x)-d+1)       % Last few frames are discarded.
        b=l:l+d-1;
        inst_freq(1,j) = inst_freq_newmethod_complex(CTab(b),d);
        ss(1,j)=sum(hab(b))/d;
        j=j+1;
    end

    %S(i,:)=ss;                     % This is the nerve spike ...
        density

    M=inst_freq.*ss;
    SN(i,:)=log(abs(diff(M)));
    %% Nonlinearity, it can be either be log/cube root/ any ...
        other nonlinearty

    %SN(i,:)=log(ss);
    clear c1 c2 c3 Si Tab hab M
    end


    %%Discrete Cosine Transform

    Ydct=dct(SN);
    cfcc=Ydct(:,2:Nc+1);

    [¬,nanC]  = find(any(isnan(cfcc)));
    cfcc(:,nanC) = [];
```

```
    [¬,infC] = find(any(isinf(cfcc)));
    cfcc(:,infC) = [];

%     ∆=∆s(cfcc,3);
%     double_∆=∆s(∆,3);
%     cfcc = [cfcc;∆;double_∆];
%     cfcc = cmvn(cfcc, 'true');
    feat = cfcc;

end
```

MATLAB code of IF estimation from QESA

```
function freq  =  inst_freq_newmethod_complex(x,c)
% zero = [0];
% y = [zero;x];
y=x;
d= diff(y);
num = [(1- (ESA(real(d))+ESA(imag(d)))) 0];
den=(2*ESA(real(x))+ESA(imag(x)));
arg=num./den;
freq = arg;
freq = sum(abs(acos(arg)))/c;
end
```

MATLAB code of ESA algorithm

```
function ESA = si(x)

for i=2:length(x)-1
    si(i-1) = x(i)*x(i)-x(i-1)*x(i+1);
end
ESA = si;
end
```