Design of Spoof Speech Detection System: Teager Energy-Based Approach

by

Madhu R. Kamble
201521005

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

DOCTOR OF PHILOSOPHY
to
DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

February 2021

Declaration

I hereby declare that

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

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

Madhu R. Kamble
(ID: 201521005)

Certificate

This is to certify that the thesis work entitled, "*Design of Spoof Speech Detection System: Teager Energy-Based Approach*," has been carried out by *Madhu R. Kamble* (201521005) for the degree of *Doctor of Philosophy* at *Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar* under my supervision.

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

Acknowledgments

First and foremost, I would like to express my gratitude and acknowledge to my Ph.D. supervisor, Prof. (Dr.) Hemant A. Patil. He constantly motivates and encourages me with helpful discussions for good quality research followed by publications in top journals and conferences. He is completely dedicated towards his students with respect to manuscript and thesis corrections, research meetings, and timely feedback on the research works. Prof. Patil has been supportive and has given me the wide freedom to pursue explore the research without objection. I am also very thankful to him for providing me opportunities for presenting my research work in various summer schools at DA-IICT, IEEE-WIE 2016, APSIPA-ASC 2017, and Nirma University. I also thank him for being there to support me during my financial problems and also he supported me whenever I was ill. In addition, I also thank his family members, Ma'am and his two kids with whom every year I enjoyed the Kite festival (Sankranti) which is good break once a while and also thank Ma'am for inviting all the Speech Lab members for the Birthday celebration which make us feel homely.

I also have to thank the members of my Ph.D. committee Prof. M. V. Joshi, Prof. Yash Vasavada, Prof. Rajib L. Das and Prof. Shruti Bhilare. I acknowledge my thesis examiners, namely, Prof. Tomi H. Kinnunen (University of Eastern Finland) and Prof. Sriram Ganapathy (IISc, Bangalore) for their excellent thesis reviews, corrections, and technical suggestions. I also acknowledge Shri Soman Nair (Executive Registrar) and Prof. Suman Mitra (Dean Academic Programs at DA-IICT) for continuously supporting me for Rajiv Gnadhi National Fellowship (RGNF) for my doctoral studies. I am also thankful to anonymous reviewers of our research papers for their valuable feedback and suggestions. I also acknowledge the funding and support from the University Grants Commission (UGC), Delhi, India and also the authorities of DA-IICT, Gandhinagar, India. In addition, I also thank IEEE Signal Processing Society (SPS) and ISCA for providing the student travel grants to present my co-authored papers during ICASSP 2019, Brighton, London, UK and INTERSPEECH 2017, Stockholm, Sweden. Furthermore, I thank all the organizers of ASVspoof 2015, 2017, and 2019 challenge, because of whom I could have few publications on the publicly available corpora.

I also thank Prof. Haizhou Li, it was a great pleasure and honor to offer a joint tutorial with Prof. Patil and Prof. Li during APSIPA-ASC 2017. In addition, it was great honor to co-author survey research article in APSIPA Transaction on SIP. I also thank Prof. Li for financial support for Article Processing Charges (APC) of 1000 US Dollars in APSIPA Transaction on SIP and paper registration ISCSLP 2018, Taipei, Taiwan. I also thank Prof. Hema A. Murthy, she was an external examiner for my JRF to SRF conversion exam. Thank you for your valuable time and suggestions on my earlier studies. I also thank Dr. Vikram Vij, Vice President of Voice Intelligence Group, Samsung Research Institute-Bangalore (SRI-B) for providing the student internship at SRI-B during the period of May-Nov, 2019. During the internship, I learn many new research domains, such as Automatic Speech Recognition (ASR), and Voice Assistants (VAs). This internship also gave me financial support.

My heartfelt gratitude and thanks to all the members of Speech Research Lab at DA-IICT. The technical discussions with Dr. Maulik and Dr. Hardik are really helpful to me for thesis and understanding speech signal processing. Special thanks to Dr. Nirmesh who always made me realize the reality and supported me in my bad times. Dr. Hardik and Dr. Nirmesh always took care of me like a brother and hence, there were always healthy fights between sister and brothers. I acknowledge my co-authors, Hemlata, Siva, Aditya, Ankur, Rajul, Gauri, and Kuldeep for exploiting the Teager Energy-based features in their research areas followed by joint publications. I also thank Dr. Anshu, Dr. Tanvina, Dharmesh, Rishabh, Prasad, Neil, Sreeraj, Srinivas, Divyesh, Mirali, Priyanka, Dipesh, Siddhant, Srishti, Harshit, Mihir, Mehak, Hetal, Falguni for having some or the other way round few memories during my Ph.D. If I have forgotten anyone, I apologize.

I thank my Ph.D. colleagues Sujata, Purvi, Rishikant, Rahul, Pankaj, Parth, Kamal, Sumukh, Gignesh, Prashant, Nidhi, Archana, and others for having hours of gossips in the cafeteria which is also equally important along with research studies for having you back into the normal life. There are many other friends and colleagues who supported me and made my Ph.D. days enjoyable at DA-IICT. I also thank my school and college friends, Akshay, Amit, Shareef, Balaji, Arvind, Kajal, Akash, Nikita (too many to list here but you know who you are!) for been there for me whenever I felt low or stress. They always make me realize that life is not only restricted to Ph.D. rather its much beyond it and so I could see different colors too.

Last but not the least I especially thank my mom, dad, and brother. My hardworking parents have sacrificed their lives for my brother and myself and provided unconditional love and care. I love them so much, and I would not have made it this far without them. I thank my brother for all his advice and support. I know I always have my family to count on when times are rough. I also thank my Grandparents and extended family members who never stopped me for pursuing my higher studies and always kept me motivated. I also thank my beloved cats, Thomya and Dumya who have both passed on. I still miss them I wished she could have lived for another few years atleast for my graduation. Things change and friends leave life doesn't stop for anybody. Now we have new edition of kittens in our family, Tokyo and Kyoto.

Contents

Abstract

Automatic Speaker Verification (ASV) systems are vulnerable to various spoofing attacks, namely, Speech Synthesis (SS), Voice Conversion (VC), Replay, and Impersonation. The study of spoofing countermeasures has become increasingly important and is currently a critical area of research, which is the principal objective of this thesis. With the development of Neural Network-based techniques, in particular, for machine generated spoof speech signals, the performance of Spoof Speech Detection (SSD) system will be further challenging. To encourage the development of countermeasures that are based on signal processing techniques or neural network-based features for SSD task, a standardized dataset was provided by the organizers of ASVspoof challenge campaigns during 2015, 2017, and 2019.

The front-end features extracted from the speech signal has a huge impact in the field of signal processing applications. The goal of feature extraction is to estimate the meaningful information directly from the speech signal that can be helpful to the pattern classifier, speech, speaker, emotion recognition, etc. Among various spoofing attacks, speech synthesis, voice conversion, and replay attacks have been identified as the most effective and accessible forms of spoofing. Accordingly, this thesis investigates and develops a framework to extract the discriminative features to deflect these three spoofing attacks.

The main contribution of the thesis is to propose various feature sets as front-end countermeasures for SSD task using a traditional Gaussian Mixture Model (GMM)-based classification system. The feature sets are based on Teager Energy Operator (TEO) and Energy Separation Algorithm (ESA), namely, Teager Energy Cepstral Coefficients (TECC), Energy Separation Algorithm Instantaneous Frequency Cepstral Coefficients (ESA-IFCC), Energy Separation Algorithm Instantaneous Amplitude Cepstral Coefficients (ESA-IACC), Amplitude Weighted Frequency Cepstral Coefficients (AWFCC), Gabor Teager Filterbank (GTFB). The motivation behind using TEO is its nonlinear speech production property. The true total source energy is known to be estimated using TEO, and it also preserves the amplitude and frequency modulation of a resonant signal and hence, it improves the time-frequency resolution along with improving the formant information rep-

resentation. In addition, the TEO also has the noise suppression property and it attempts to remove the distortion caused by noise signal.

In Chapter 3, we analyze the replay speech signal in terms of reverberation that occurs during recording of the speech signal. The reverberation introduces delay and change in amplitude producing close copies of speech signal which significantly influences the replay components. To that effect, we propose to exploit the capabilities of Teager Energy Operator (TEO) to estimate running estimate of subband energies for replay *vs.* genuine signal. We have used linearly-spaced Gabor filterbank to obtain narrowband filtered signal. The TEO has the property to track the instantaneous changes of a signal. In Chapter 4, we propose Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) features using Energy Separation Algorithm (ESA). The speech signal is passed through bandpass filters in order to obtain narrowband components because speech is a combination of several monocomponent signals. To obtain a narrowband filtered signal, we have used linearly-spaced Butterworth and Gabor filterbank. The instantaneous modulations helps to understand the local characteristics of a non-stationary signal. These IA and IF components are able to capture the information present in a slowly-varying amplitude envelope and fast-varying frequency. The slow-varying temporal modulations for replay speech have the distorted amplitude envelope, and the fast-varying temporal modulation do not preserve the harmonic structure compared to the natural speech signal. For replay speech signal, the intermediate device characteristics and acoustic environment distorts the spectral energy compared to the natural speech energy. In Chapter 5, we extend our earlier work with the generalized TEO, i.e., by varying the samples of past and future instants with a constant arbitrary integer *k* also known as *lag parameter or dependency index*, and named it as Variable length Teager Energy Operator (VTEO). In Chapter 6, we propose the combination of Amplitude Modulation and Frequency Modulation (AM-FM) features for replay Spoof Speech Detection (SSD) task. The AM components are known to be affected by noise (in this case, due to replay mechanism). In particular, we explore this damage in AM component to corresponding Instantaneous Frequency (IF) for SSD task. Thus, the novelty of proposed Amplitude Weighted Frequency Cepstral Coefficients (AWFCC) feature set lies in using frequency components along with squared weighted amplitude components that are degraded due to replay noise. The AWFCC features contains the information of both AM and FM components together and hence, gave discriminatory information in the spectral characteristics.

The first motivation in this thesis is to develop various countermeasures for

SSD task. The experimental results on the standard spoofing database shows that proposed feature sets perform better than the corresponding baseline systems. Inspired by the success in the SSD task, we applied TEO-based feature set in a variety of speech and audio processing applications, namely, Automatic Speech Recognition (ASR), Acoustic Scene Sound Classification (ASC), Voice Assistant (VA), and Whisper Speech Detection (WSD). In all these applications, our TEO-based feature set gave consistently better performance compared to their respective baselines.

List of Acronyms

AI	Artificial Intelligence
AM	Amplitude Modulation
ANN	Artificial Neural Networks
ASC	Acoustic Scene Classification
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
AWFCC	Amplitude Weighted Frequency Cepstral Coefficients
BLSTM	Bidirectional Long Short-Term Memory
BM	Basilar Membrane
CD-HMM	Continuous Density Hidden Markov Models
CF	Center Frequencies
CFCC	Cochlear Filter Cepstral Coefficients
CFCCIF	Cochlear Filter Cepstral Coefficients Plus Instantaneous Frequency
CMN	Cepstral Mean Normalization
CNN	Convolutional Neural Networks
ConvRBM	Convolutional Restricted Boltzmann Machine
CQCC	Constant Q Cepstral Coefficients
D	Dimension of a Feature Vector
DCT	Discrete Cosine Transform
DET	Detection Error Trade-off
DNN	Deep Neural Network
DST	Deep Scattering Transform
EER	Equal Error Rate
EM	Expectation Maximization
ERB	Equivalent Rectangular Bandwidth
ESA	Energy Separation Algorithm

FAR	False Acceptance Rate
FM	Frequency Modulation
FRR	False Rejection Rate
FT	Fourier Transform
GDF	Group Delay Function
GMM	Gaussian Mixture Model
HT	Hilbert Transform
HMM	Hidden Markov Model
IA	Instantaneous Amplitude
IACC	Instantaneous Amplitude Cepstral Coefficients
IFCC	Instantaneous Frequency Cepstral Coefficients
ISO	International Organization for Standardization
LDA	Linear Discriminant Analysis
LFCC	Linear Frequency Cepstral Coefficients
LLR	Log-Likelihood Ratio
LM	Language Model
LSTM	Long Short-Term Memory
LTI	Linear Time-Invariant
MBR	Minimum Bayes Risk
MFCC	Mel Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
RNN	Recurrent Neural Networks
SAS	Spoofing and Anti-Spoofing
SGD	Stochastic Gradient Descent
SS	Synthetic Speech
SV	Speaker Verification
SVM	Support Vector Machine
SSD	Spoof Speech Detection
STFT	Short-Time Fourier Transform
TDNN	Time-Delay Neural Networks
TEO	Teager Energy Operator
TECC	Teager Energy Cepstral Coefficients
TFS	Temporal Fine Structure
TTS	Text-to-Speech Synthesis

USS	Unit Selection Synthesis
VC	Voice Conversion
VA	Voice Assistant
VESA	Variable Length Energy Separation Algorithm
VTEO	Variable Length Teager Energy Operator
WER	Word Error Rate
WSD	Whisper Speech Detection
WSR	Whisper Speech Recognition

List of Symbols

t	Time
ω	Frequency
$x(t)$	Speech Signal
$x(n)$	Discrete-time Speech Signal
$x_a(t)$	Analytic Signal
$g(t)$	Impulse Response of the Gabor Filter
$G(z)$	System Function in Z-domain
$G(e^{j\omega})$	Frequency Response of System
$\Psi\{\cdot\}$	Teager Energy Operator
$a_i[n]$	Instantaneous Amplitude
$\Omega_i[n]$	Instantaneous Frequency
$E(\cdot)$	Energy Function of an Energy-Based Models
$\sum$	Notation for Summation
θ	Model Parameters
$\mathbf{x}$	Input Signal or Visible Units to the Learning Algorithm
$\mathbf{X}$	Feature Vectors
$p(\mathbf{X}\|H_0)$	Likelihood Scores
α	Fusion Parameter
$\phi(t)$	Instantaneous Phase
$*$	Convolution Operation
$\oplus$	System Combination
$\odot$	Elementwise Multiplication
$\approx$	Approximately
Δ	Delta or Dynamic Features
$\Delta\Delta$	Delta-Delta or Double Delta or Acceleration Features
$\mathbb{E}[\cdot]$	Expectation Operator

C^∞	Space of Infinitely Differentiable Functions
$L^2(R)$	Hilbert Space of Square Integrable Functions
$arcsin$	Inverse Sine Function
$arctan$	Inverse Tangent Function
$H\{\cdot\}$	Hilbert Transform Operator
$Rxx(\tau)$	Autocorrelation Function for $x(t)$ with Lag τ

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Motivation

One expects voice to be the primary source of interface between humans, and machines in the near future [32]. Various biometric traits that have been successfully used in practical applications include voice, face, iris, fingerprint, palmprint, palm/finger vein, etc. [33]. The Automatic Speaker Verification (ASV) system is a biometric speaker authentication to verify a claimed speaker's identity with the help of the machines [34]. According to major companies that are involved in the speech recognition research believe that the perfect user interface, does not exist till date and to build it, knowledge of both sociology and technology fields are required [35]. The developed systems allows, one to wirelessly control lights, fans, TV, AC, security, etc [36]. Home automation now-a-days is one of the major growing industries that changes the lifestyle of people [35]. On the other hand, for special need, such as the person who are elderly and the disabled [37]. It is also useful for the people who live alone might require helping hand at home [37].

Though the ASV systems are convenient and easy to use, it raise new security issues because of their vulnerability to several types of spoofing attacks, such as replay, impersonation, synthetic speech, voice conversion [1, 38]. In practice, we would like an ASV system to be robust against variations, such as microphone and transmission channel, intersession, acoustic noise, speaker aging, etc. This robustness makes an ASV system to be vulnerable to various spoofing attacks as it tries to *nullify* these effects. The spoofing attacks in biometrics are also known as *presentation attacks* as per International Organization for Standardization (ISO), and International Electro-technical Commission (IEC) [39].

Impersonation is defined as the process of producing the similar voice pattern, and speech behavior of the target speaker's voice [40–42]. The impersonators do not require any technical background knowledge or machines to imitate the target speaker. Speech synthesis research is also carried out using Text-To-Speech

(TTS) system, where the text is given as the input, and the system produces a speech signal at the output [43–45]. It is a machine-generated voice production system, and represents a genuine threat. Voice Conversion (VC) is the process of converting the source speaker's voice to a sound similar to that of target speaker's voice [41, 46, 47]. One of the most easiest and simple spoofing attack is the replay attack. The replay is a pre-recorded speech signal of a target speaker's voice that is captured using a recording device to get the fraudulent access to the system [48–50].

Due to recent technological developments, it is possible to generate spoofed data that resembles very close to their natural counterparts. To protect the ASV system from such attacks, one of the approach is to develop an ASV system that is more robust to resist any kind of spoofing attack. Another approach is to build an independent Spoof Speech Detection (SSD) system. These spoofed signals resemble close to the corresponding natural voice and thus, have high chance to spoof the ASV system. The illustration of spoofing on ASV system is shown in Figure 1.1. It is therefore very important to develop countermeasures that can detect such spoofing attacks.

Specifically, voice-based access in digital devices, such as smartphones, is equipped with an AI-enabled personal assistant, e.g., OK Google, Apple Siri, Microsoft Cortana, Amazon Echo, etc. Since an AI-enabled framework is used in real-life noise scenarios, we need a knowledge of both signal processing and machine learning to propose a feasible solution to develop a noise-robust voice-based access system. Hence, there is a great demand for speech and audio technology in the near future, since voice is the most important form of human communication, and possibly human-to-machine interactions.

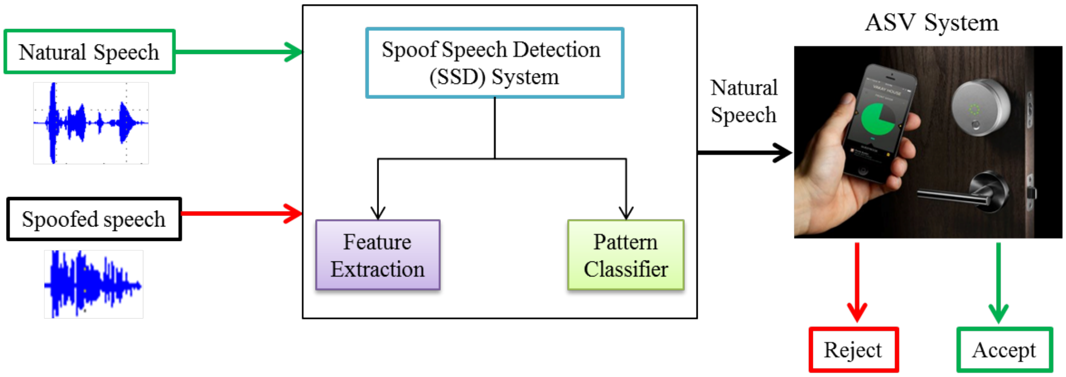


Figure 1.1: Basic Illustration of Spoofing on ASV System. After [18].

The objective of this thesis is the study of Energy Separation Algorithm (ESA), and Teager Energy Operator (TEO)-based features for SSD task. The motivation behind using TEO is its nonlinear speech production property [51]. The true total source energy is estimated using TEO, and it also preserves the amplitude and frequency modulation of a resonant signal and hence, it improves the time-frequency resolution along with improving the formant information representation [52]. In addition, the TEO has the noise suppression capability and thus, it attempts to suppress the distortion caused by additive noise signal.

The Amplitude Modulation (AM) and Frequency Modulation (FM)-based features have been used for several speech applications (involving signal degradation), such as speech and speaker recognition, speech analysis, and synthesis, etc. In particular, modulation features from the AM-FM speech model were originally conceived for robust speech recognition task [53] as they capture the second order non-linear structure of speech formants. Very recently, these features are applied for far-field/whisper speech recognition (WSR) task [54–56]. In this study, we explore the significance of AM and FM representation for replay SSD task. Recently, we have proposed AM-FM-based feature sets using demodulation techniques with Hilbert Transform (HT), and Energy Separation Algorithm (ESA) [27]. The motivation behind using the TEO and ESA approach is as follows:

- TEO has the capability to capture property of airflow pattern in vocal tract system during natural speech production and thus, it can be an excellent use for SSD task.

- The ESA is applied on the narrowband filtered speech signals, which are modeled using AM-FM signals to estimate time-varying amplitude envelope, and instantaneous frequencies [57].

- The ESA approach do not require the computationally complex task of phase unwrapping (as it is required for HT-based approach of analytic signal generation).

- To estimate the IA and IF components with ESA approach, only five consecutive samples are required.

- The slow and fast-varying temporal modulations obtained at different time scale have the distortion for the replay speech signal compared to the natural speech.

- The IF component estimated for the subband filtered signal shows the damping in the fluctuation for the replay signal around the center (carrier) frequency.

- For the same time scale from where the IF fluctuation started having tilt from its center frequency, sinc-like patterns are observed in replay signal than its natural counterpart in the voiced regions.
- The spectral energy obtained from the Teager Energy Operator (TEO), shows the difference for the natural, and its corresponding replay speech signal in all the frequency regions, which is not captured by the traditional spectrogram.

Our proposed feature sets has been applied for various applications as given below:

- Spoof Speech Detection (SSD) [1,9–18,27–29,58–60].
- Replay Detection for Voice Assistants (VAs) [20,21].
- Automatic Speech Recognition (ASR) [19,61].
- Whisper Speech Detection (WSD) or classification of normal *vs.* whisper speech [22].
- Acoustic Scene Classification (ASC) [23].

1.2 Key Research Challenges

The key research challenges in developing the SSD system are as follows:

- Performance of Joint Protocol with ASV Systems: The current studies on developing countermeasures, and ASV systems are carried out independently. What user would like to have is a secure and accurate ASV system. However, a more robust ASV system to noise and channel variations may become less secure against spoofing attacks. As there is no guarantee of having a better performing countermeasure that provides lower EER, and also reliable for the ASV system performance. Hence, with the progress made in the research of spoofing detection, evaluation metrics must evolve to reflect the joint protocol system performance.
- Liveness Detection: The use of high quality recording loudspeaker or playback device to record/playback the speech signal. In this process, the quality of signal captured becomes indistinguishable from live human voice. This high quality device makes the spoofed speech signal almost impossible to detect that depends on the acoustic cues. This gives motivation to investigate further on the liveness detection of human voice.

- Logical and Physical Access: The physical access is the actual spoofing, where the speech is played back through a microphone into the ASV system. However, the ASVspoof database gave a special attention to logical access attacks. For such attacks, it is assumed that the spoofed samples are directly injected into the system through a software-based process [62]. Hence, physical access attacks might be more realistic than the logical access attacks, where the attacker plays back a recorded utterance to the system. This utterance can be either obtained from the real speaker or can be forged using voice conversion (VC) or synthetic speech (SS) algorithms. This motivates the study on physical access attacks, and evaluation database development.

- Comparison of Human *vs.* Machine Learning:
  It is of great interest to know whether human perception is important in identifying spoofing and hence, humans can achieve better performance than automatic approaches in detecting spoofing attacks. There was a benchmark study comparing automatic systems against human performance on a speaker verification and synthetic speech spoofing detection tasks (synthetic speech and voice conversion spoofs) [63].

- Robustness to High Quality Speech Synthesizers:
  Recently, many representation learning-based high quality speech synthesis techniques were proposed that achieved significantly better naturalness. The Wavenet [64], GAN [65], and other end-to-end speech synthesis architectures produce high quality synthesized speech [66]. It is also shown that low quality publicly available database can be used to produce high quality spoof data using GAN-based speech enhancement [67]. Such high quality synthetic speech and voice conversion techniques may further increase the difficulties in synthetic SSD. This technique could be used to generate spoof speech database in the next edition of ASVspoof challenge [7].

- Robustness to Signal Degradation or Noisy Conditions:
  Current publicly available spoofing databases are developed in clean conditions. However, the recent replay database was recorded under various acoustic environmental conditions. For ASVspoof 2015 challenge database, the noisy database was developed by adding various noises at different Signal-to-Noise Ratio (SNR) levels. Further investigations are required as to how the diversity of different noise types affects the SSD performance. In addition, the study is required to observe the effect on SSD, when the additive noise is added manually, and when the noise is added naturally

via the acoustic environment. Study for different acoustical background, microphone, etc. is reported in [6]. Hence, the countermeasures must be developed that it should be robust to signal degradation conditions as well.

1.3 Contributions from the Thesis

The main contribution of this thesis is to propose various feature sets as front-end countermeasures for SSD task. The feature sets are based on TEO and ESA speech demodulation technique. Figure 1.2 shows the key features sets proposed in the thesis. Following are the key contributions in this thesis.

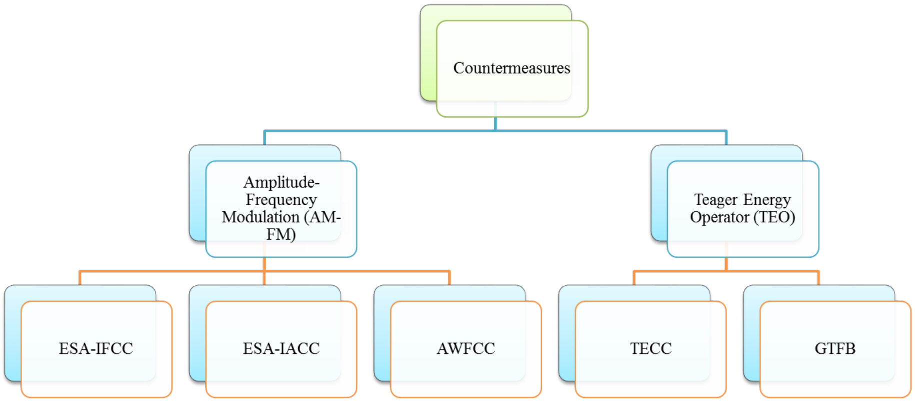

Countermeasures
Amplitude-Frequency Modulation (AM-FM)
Teager Energy Operator (TEO)
ESA-IFCC
ESA-IACC
AWFCC
TECC
GTFB

Figure 1.2: Various Countermeasures Proposed in This Thesis. ESA- Energy Separation Algorithm, IF/IACC- Instantaneous Frequency/Amplitude Cesptral Coefficients, AWFCC- Amplitude Weighted Frequency Cepstral Coefficients, TECC- Teager Energy Cepstral Coefficients. GTFB- Gabor Teager Filterbank.

1.3.1 Proposed Feature Sets for SSD

Recent research focuses on either to improve the novel features or improving the back-end modeling for SSD task. The features extracted are supposed to deal with speaker variability, phonetic variability, channel effects, acoustical environment, etc. The replay speech is mainly distorted because of background noise (such as, air-conditioner, computer), echo, and reverberation. Replay speech gets distorted by both interfering sounds and reverberation caused because of the recording of target speaker's voice from the distance. The reverberation introduces delay and change in amplitude producing close copies of speech signal, which significantly influences the replay components. To that effect, we propose to exploit the capabilities of Teager Energy Operator (TEO) to estimate running estimate of subband energies for replay *vs.* genuine signal. We have used linearly-spaced Gabor filterbank in order to obtain narrowband filtered signal. The TEO has the capability to track the instantaneous changes of a signal.

1.3.2 Speech Parametrization (Analysis) of the Feature Sets

We consider the effect of different frequency scales, type of the filterbank used, and number of subband filters used in filterbank. In order to emphasize the higher frequency regions, the pre-emphasized speech signal is used as an input to the filterbank. For the natural speech, as we locate the center frequency to higher formants the energy traces are found to have more *bumps* than the energy traces obtained from the lower formants. However, these energy traces were not found for the replay speech signal as we locate the center frequency towards higher formants. The variations in TEO profiles were observed for the replay speech signal, when recorded in different acoustical environments. The slow and fast-varying temporal modulations obtained at different time scale have the distortion for the replay speech signal compared to the natural speech. The spectral energy obtained from the TEO, shows the difference for the natural, and its corresponding replay speech signal in all the frequency regions, which is not captured by the traditional spectrogram.

1.3.3 Applications of Teager Energy-based Features

The first motivation to develop various feature sets as countermeasures for Spoof Speech Detection (SSD) task. The experimental results on the standard SSD datasets indicate that our proposed ESA-based features perform better than the acoustic feature sets along with the baseline system. Later, the TEO-based feature set is applied for a variety of speech and audio processing applications, namely, Automatic Speech Recognition (ASR), Voice Assistants (VAs), Whisper Speech Detection (WSD), and Acoustic Scene Classification (ASC). In all these applications, our proposed feature set gave consistently better performance compared to the respective baselines. The overall contributions of this thesis are summarized in Figure 1.3.

1.4 Organization of the Thesis

The organization of the rest of the chapters in the thesis is shown via a flowchart in Figure 1.4 and briefly described next.

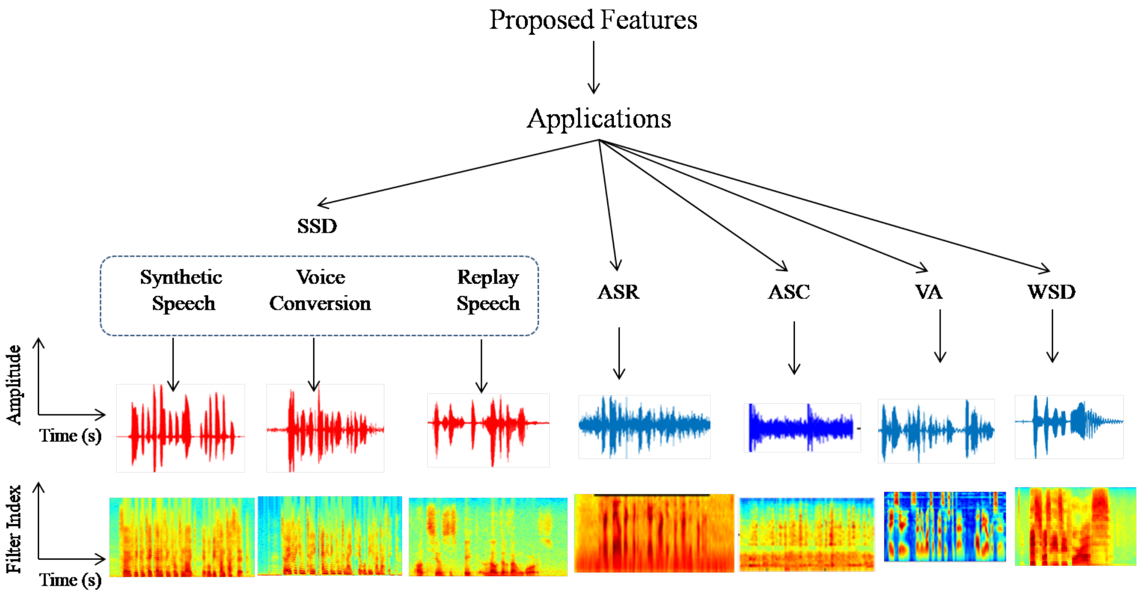


Figure 1.3: Pictorial Representation of Proposed Features Applied in Different Applications. ASR -Automatic Speech Recognition, ASC= Acoustic Scene Classification, VA- Voice Assistant, WSD- Whsiper Speech Detection.

- **Chapter 2** discusses the literature search on spoofing attacks for voice biometrics. Different types of spoofing attacks are discussed with focus on replay attacks. Various approaches or methods focusing on the traditional and Representation Learning (RL) approaches are also discussed. The several issues with the stand-alone detectors are also briefly discussed.

- **Chapter 3** discusses the detailed analysis to understand modeling of replay signal, reverberation mechanism, and relevant basics of Teager Energy Operator (TEO). The reverberation introduces delay in the genuine speech components corresponding to different transmissions and reflections that further depends on the acoustical environmental conditions. The experiments are performed on different databases used for SSD task to evaluate the performance of proposed feature set.

- **Chapter 4** discusses the proposed feature sets used for SSD task, namely, Energy Separation Algorithm Instantaneous Frequency Cepstral Coefficients (ESA-IFCC), and Energy Separation Algorithm Instantaneous Amplitude Cepstral Coefficients (ESA-IACC). The proposed feature sets are based on Energy Separation Algorithm (ESA), and Teager Energy Operator (TEO). The experiments on the SSD task using the standard datasets are presented to evaluate the proposed feature set.

- **Chapter 5** discusses our earlier work extended with the generalized TEO,

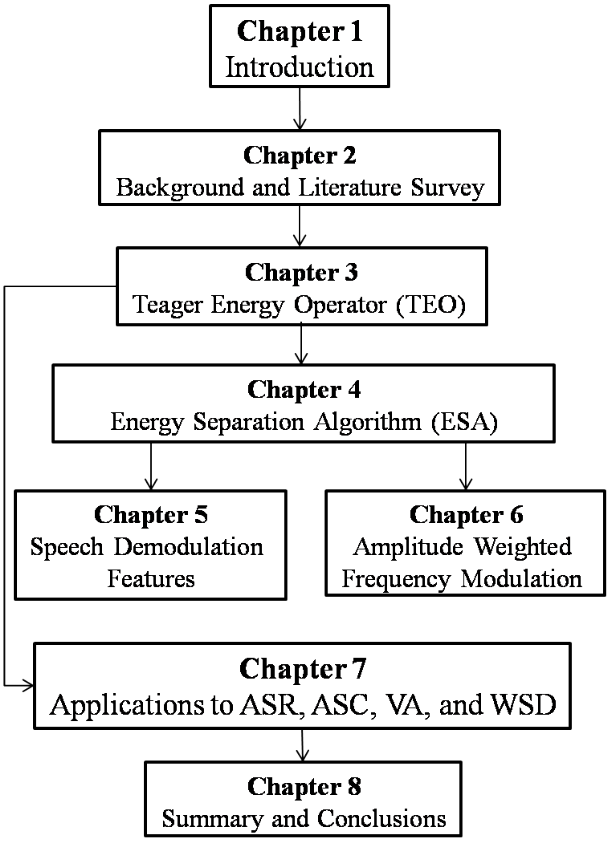

Chapter 1
Introduction
Chapter 2
Background and Literature Survey
Chapter 3
Teager Energy Operator (TEO)
Chapter 4
Energy Separation Algorithm (ESA)
Chapter 5
Speech Demodulation Features
Chapter 6
Amplitude Weighted Frequency Modulation
Chapter 7
Applications to ASR, ASC, VA, and WSD
Chapter 8
Summary and Conclusions

Figure 1.4: Flowchart Depicting Organization of this Thesis.

i.e., by varying the samples of past and future signal with a constant arbitrary integer *k* also known as *lag parameter*, and named it as Variable length Teager Energy Operator (VTEO). We compared the Variable length Energy Separation Algorithm (VESA)-based features with earlier proposed method, i.e., ESA along with Hilbert transform method for SSD task. In particular, we performed experiments on ASVspoof 2017 Version 2.0 challenge database, and BTAS database.

- **Chapter 6** discusses importance of using the combined information of AM and FM components rather than using it alone for replay SSD task. The AM and FM components are estimated via the Energy Separation Algorithm (ESA) that works on narrowband subband filtered signals. The Instantaneous Amplitude (IA) obtained from the Amplitude Modulation (AM) component of a narrowband speech signal is severely affected by the noise and multipath interference (such as, due to replay mechanism). The noise present in the replayed signals are explored by the IF components. In particular, this damage in AM components is exploited by the proposed feature set.

- **Chapter 7** discusses the Teager energy-based features used for different application, namely, Automatic Speech Recognition (ASR), Acoustic Scene Classification (ASC), and Voice Controlled Systems (VCS) along with the Whisper Speech Detection (WSD). The experiments are performed on standard dataset, and found the Teager energy-based feature set perform better compared to the corresponding baseline systems.

- **Chapter 8** concludes and summarizes the work done in the thesis. The contributions in the thesis are presented. The Chapter also discusses the applications, limitations of the present work, and future research directions for the task of anti-spoofing presented in the thesis.

1.5 Chapter Summary

This Chapter gave an outline of the basic ASV system and introduced the problem of anti-spoofing. The motivation and need of anti-spoofing measures (or countermeasures for spoofing) for our research work in this thesis is discussed. The major contributions in the thesis include a novel feature sets, analysis of the feature sets, and applications to SSD, ASR, ASC, VCS, and WSD. The organization of various chapters in this thesis is also presented. In the next Chapter, we discuss the background studies, and the literature corresponding to spoofing attacks along with the limitations in the literature, and current research issues in order to understand gap area for research and development for anti-spoofing for ASV.

CHAPTER 2

Background on Spoofing Attacks, Databases, Countermeasures

2.1 Introduction

This Chapter discusses the literature search on the Spoofed Speech Detection (SSD) task [68]. In the literature [2], the spoofing attacks are broadly classified into four types, namely, impersonation, synthetic speech (SS), voice conversion (VC), and replay. The detailed description of each spoofing attack along with their generation method, availability, and risk factor to be considered, when dealing with the SSD task is discussed next. Development of the vulnerability of Automatic Speaker Verification (ASV) system to speech synthesis, voice conversion, and replay attacks is presented. The anti-spoofing countermeasures existing in the literature for both machine-generated, and replay speech signals are discussed. This Chapter brings out briefly various research issues (i.e., gap area in SSD field) for the SSD task, majority of which will be addressed in this thesis work.

We have seen a surge of research papers on spoofing detection in scientific conferences, such as APSIPA Annual Summit and Conference [24], ICASSP, INTERSPEECH, and special issues in scientific journals as reported in Table 2.1.

2.2 Spoofing Attacks

The spoofed speech can be generated by humans themselves or by using different machine-based approaches. The impostor speech generated by humans will be classified under impersonation (or mimicking/identical twins). On the other hand, impostor developed using artificially using machines can be due to replay, synthetic speech, and voice conversion techniques. A brief discussion on these attacks in voice biometrics is discussed in this sub-Section along with few spoofing algorithms as shown in Fig. 2.1.

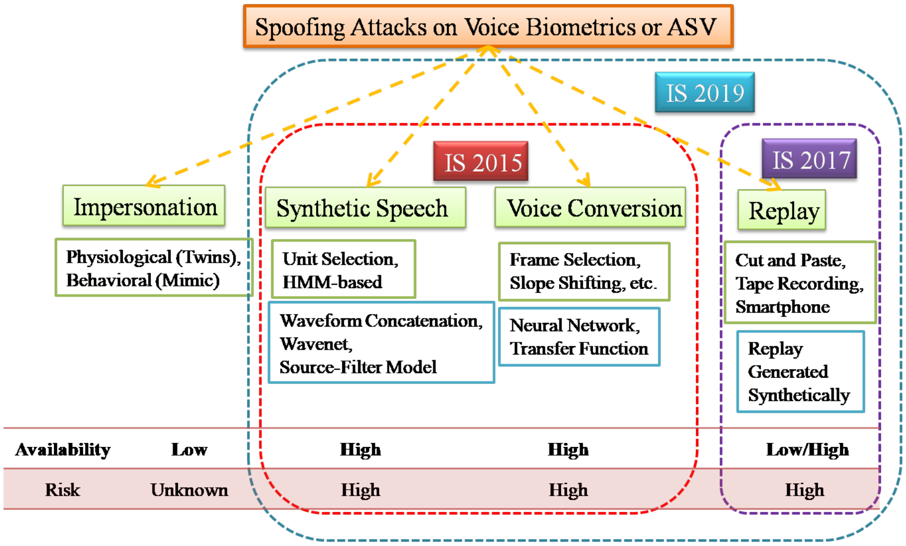

Spoofing Attacks on Voice Biometrics or ASV
IS 2019
IS 2015
IS 2017
Impersonation
Physiological (Twins), Behavioral (Mimic)
Synthetic Speech
Unit Selection, HMM-based
Waveform Concatenation, Wavenet, Source-Filter Model
Voice Conversion
Frame Selection, Slope Shifting, etc.
Neural Network, Transfer Function
Replay
Cut and Paste, Tape Recording, Smartphone
Replay Generated Synthetically
Availability
Low
High
High
Low/High
Risk
Unknown
High
High
High

Figure 2.1: Different Spoofing Attacks on Voice Biometrics Along With Their Availability, and Risk Factor. IS: INTERSPEECH. Adapted from [1,24].

Table 2.1: List of Special Issues on ASVspoofing and Countermeasures. After [1]

No.	Details of Special Issue
1	IEEE Transactions of Information Forensics special issues on Biometrics Spoofing and Countermeasures [69]
2	IEEE Signal Processing Magazine special issue on Biometric Security and Privacy Protection [70]
3	IEEE Journal on Selected Topics in Signal Processing on Spoofing and Countermeasures for Automatic Speaker Verification [71]
4	Speaker and Language Characterization and Recognition: Voice Modeling, Conversion, Synthesis and Ethical Aspects [72]
5	Advances in Automatic Speaker Verification Anti-spoofing [73]

2.2.1 Impersonation

Impersonation is defined as the process of imitating the similar voice pattern, and speech behavior of the target speaker's voice [40–42]. This can be done by professional mimics/impersonator (by utilizing behavioral characteristics) or by twins (by utilizing physiological characteristics) or anyone can (try to) impersonate another person [74]. The impersonators do not require any technical background or the machines to imitate the target speaker. The study in [75] found that if the impostor is aware of the claimed speaker's voice, and also carries similar voice pattern could crack the biometric system. For better imitation, the professional imitator tries to mimic the prosodic features of a target speaker [76]. Professional voice imitator, intend to mimic the claimed speaker's prosody, accent, pronunciation, lexicon, and other high-level speaker traits. Such imitation may mislead human perception, however, it is less effective in attacking ASV systems because most ASV systems are based on the spectral features to make decisions. Just like twins attacks, in impersonation attacks, the system is presented with natural human speech. A system to detect unnatural speech does not help. As it takes special training to impersonate someone's voice, impersonation attack is not considered as a common threat to ASV systems.

In speaker recognition, we aim to extract the unique speaker features from the speech data. However, the speaker-specific features become relatively less unique between the twins [77]. Generally, spectrographic analysis is used to identify the speaker's voice. In the case of identical twins, the same technique fails to perform primarily due to similarity in shape and size of vocal tract system [78]. The study reported in [79], states that the pattern of speech signals, pitch (fundamental frequency, F_0) contours, formant contours, and spectrograms for identical twin speakers are very similar, if not identical. Due to lack of uniqueness, the FAR increases for ASV of identical twins. Twins attacks are also referred to as *twins fraud* [77]. Recently, the Voice ID service was launched by HSBC's phone banking business, however, it failed to recognize true speaker [80–82]. Similar twins fraud was studied in other biometrics literature as well [77], and a dedicated doctoral thesis in this area [83]. The identical twins do have a similar spectrographic pattern, however, the ASV technology has seen a significant reduction in fraud, and has proven to be more secure than ATM pins, passwords, and memorable phrases. In twins attacks, the system is presented with natural human speech, a synthetic speech detection mechanism will not enhance the security of the system. To distinguish between the twins, further study on discriminative speaker features is required or more study in this direction is required as observed more than four decades earlier in [74].

2.2.2 Synthetic Speech

Synthetic speech is also referred to as Text-To-Speech (TTS) voice, which takes text as input, and generate speech as output. It emulates a human vocal production system, and represents a genuine threat. The synthetic speech is now able to generate high quality voice due to recent advances in unit selection synthesis [43], statistical parametric [44], hybrid [45], and DNN-based TTS methods. Recently, deep learning-based techniques, such as Generative Adversarial Network (GAN) [65], Tacotron [84], Wavenet [64]. are able to produce very natural-sounding speech both in timbre, and prosody. The synthetic speech uses properties of a claimed speaker's voice characteristics, and spectral cues of the natural speech. The spectral energy density of natural (Panel I), and synthetic (Panel II) speech signal are shown in Fig. 2.2. It is clearly observed that the distributions of spectral energies are very different between the natural speech, and synthetic speech. The research on synthetic speech detection has been focused on how to detect the artifacts that exist in the synthetic speech. More technical description of algorithms are reported in [85,86].

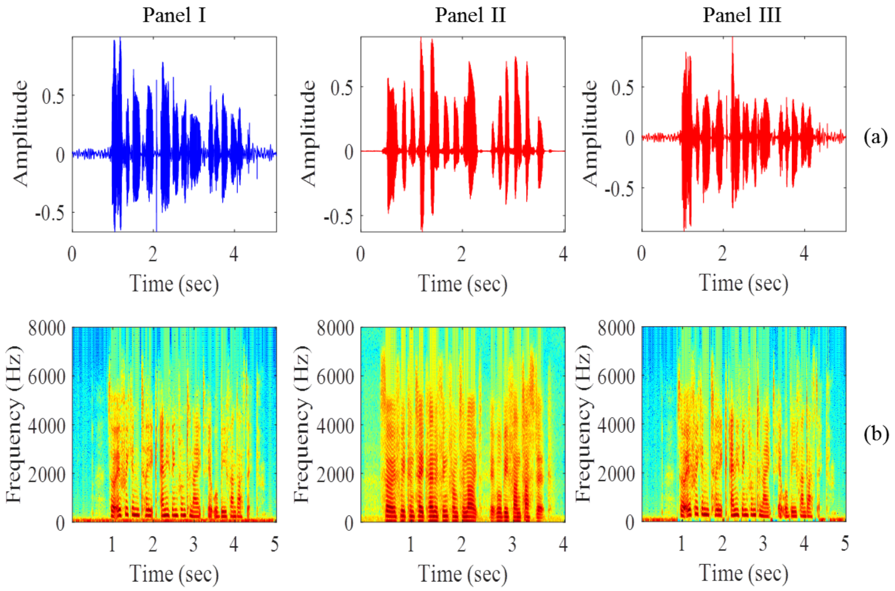


Figure 2.2: Spectral Energy Densities of Natural (Panel I), Synthetic Speech (Panel II), and Voice Converted Speech (Panel III). (a) Time-domain Speech Signal, and (b) Corresponding 2-Dimensional (2-D) Spectral Energy Density.

2.2.3 Voice Conversion

Voice Conversion (VC) is the process of converting the source speaker's voice to a sound similar to the target speaker's voice [41,46,47]. Voice conversion deals with the information that relates to the segmental and suprasegmental features, while keeping the language content similar [87]. The earlier studies includes statistical techniques, such as Gaussian Mixture Model (GMM) [88], Hidden Markov Model (HMM) [89], unit selection synthesis [90], principal component analysis (PCA) [91], and Non-negative matrix factorization (NMF) [92] for VC task. Recently, DNN [93], Wavenet [64], and GAN [65] represents a technology leap.

Studies also reported in the area of signal processing techniques, such as vector quantization (VQ) [94], and frequency warping (FW) [95]. The research on voice conversion detection has also been focused on how to detect the artifacts arising from the voice conversion process. One example of the converted speech is illustrated in Panel III of Fig. 2.2. More technical description of converted voices are reported in [85], [87].

2.2.4 Replay

One of the most accessible spoofing is replay attack. The attacker replays a pre-recorded voice from the target speaker to the system to gain access [48–50]. Such attack is meaningful for text-dependent as well as for text-independent ASV systems. With high quality record-replay audio device, the replayed speech is highly similar to the original speech, spectral content will change slightly due to impulse response of an device that is modelled as LTI system (primarily due to convolution theorem). Hence, replay is a serious adversary to text-dependent ASV system.

Fig. 2.3 shows the spectrographic analysis of natural speech, and replay speech signal taken from the ASVspoof 2017 Challenge database [5]. The Panel I of Fig. 2.3 shows the natural speech signal with the corresponding spectrogram of the natural speech signal for the utterance, *"Actions Speak Louder Than Words"*, and similarly, Panel II is for the replayed speech signal. It can be observed from the Fig. 2.3 that there is a difference in *temporal* as well as in *spectral* representation between Panel I (natural), and Panel II (replay) speech signal due to the channel, and noise distortion.

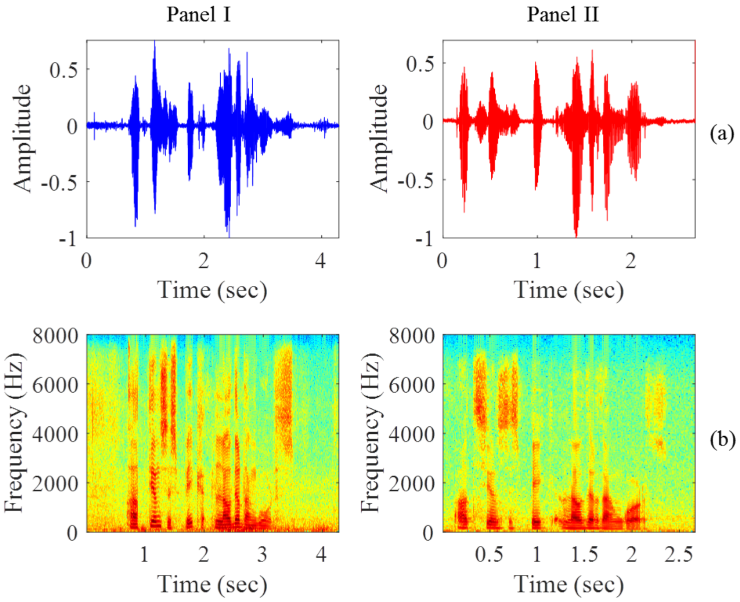


Figure 2.3: Spectral Energy Densities of Natural (Panel I), and Replay Speech (Panel II). (a) Time-Domain Speech Signal, and (b) 2-Dimensional (2-D) Spectral Energy Density.

2.3 Database for SSD

To objectively report the research progress, there is a need to provide a common dataset along with performance metric to evaluate the spoofing countermeasures. This was also discussed in the special session on spoofing and countermeasures for ASV held during INTERSPEECH 2013 [68]. This special session motivated the researchers to organize the first ASVspoof 2015 Challenge held in INTERSPEECH 2015 [2]. The database released in this challenge contains two types of spoofing attacks, namely, synthetic speech, and voice conversion. As a follow up, the second and third challenges were organized during INTERSPEECH 2017, and INTERSPEECH 2019, respectively [5], [96]. The historical developments and key milestones of the ASVspoof initiative are illustrated in Fig. 2.4.

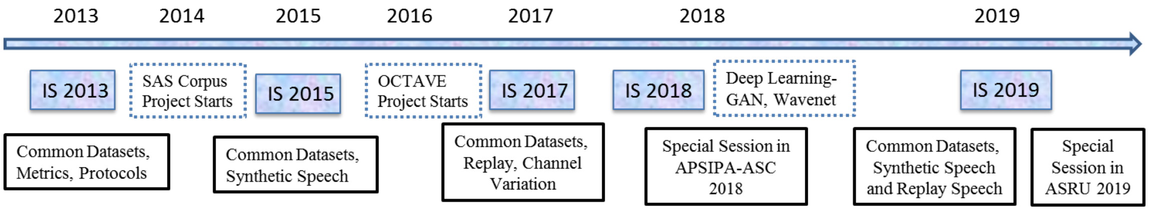

2013
2014
2015
2016
2017
2018
2019
IS 2013
SAS Corpus Project Starts
IS 2015
OCTAVE Project Starts
IS 2017
IS 2018
Deep Learning-GAN, Wavenet
IS 2019
Common Datasets, Metrics, Protocols
Common Datasets, Synthetic Speech
Common Datasets, Replay, Channel Variation
Special Session in APSIPA-ASC 2018
Common Datasets, Synthetic Speech and Replay Speech
Special Session in ASRU 2019

Figure 2.4: The Selected Chronological Progress in ASVspoof Literature for Voice Biometrics. In INTERSPEECH 2013, a Special Session was Organized and Spoofing and Anti-Spoofing (SAS) Corpus of Speech Synthesis and Voice Conversion Spoofing Data was Created. The First ASVspoof Challenge was Held in INTERSPEECH 2015. In 2015, the OCTAVE Project Started which Focused on TTS, VC and Replay Spoofing Data. Only replay attack was further carried as the Second Edition of ASVspoof Challenge in INTERSPEECH 2017. The Follow Up Third ASVspoof 2019 Challenge was on Physical, and Logical Access Attacks Going to be Held during INTERSPEECH 2019 [7]. IS indicates INTERSPEECH.

The early studies of spoofing attacks used different speech, and speaker recognition databases, such as YOHO [41, 97], NIST SRE [46, 47, 98], and WSJ [86, 99]. The databases used for anti-spoofing studies are reported in Table 2.2. Since 2015, the research community has released multiple evaluation databases, that include SAS, ASVspoof 2015 challenge, ASVspoof 2017 challenge, ASVspoof 2019 challenge, AVspoof, RedDots Replayed, and ReMASC databases. The AVspoof database introduces replay spoofing attacks along with synthetic speech (SS), and voice conversion (VC) spoofing attacks. It was designed to simulate the attacks via logical and physical access. This database was used in the BTAS 2016 Challenge [3, 100]. RedDots [4] database is developed originally for text-dependent ASV research that was re-developed from replay attacks. This database is derived from the original RedDots database under various recording, and playback conditions. However, standard impersonation database is not yet available publicly, the study reported in [97] used the YOHO database that was designed for ASV system. In this chapter, we focus on description of ASVspoof challenge datasets for the years 2015, 2017, and 2019. Next, we will discuss about challenge databases in details.

Table 2.2: Various Corpora for Spoofing Attacks on ASV System. After [1]

Types of Spoofing Attack	Spoof Corpus
Impersonation [41]	YOHO
Voice Mimicry [42]	NIST
SS [99]	WSJ
SS [86]	WSJ
VC [46]	NIST SRE 2006
VC [47]	NIST SRE 2006
VC [98]	NIST SRE 2006
VC, SS, and Artificial Spoof [101]	NIST SRE 2006
SS and VC [85]	SAS
Replay [102]	RSR2015
VC and Replay [103]	RSR2015
SS and VC [38]	ASV Spoof 2015
SS, VC, and Replay [100]	AV Spoof
Replay [4]	RedDots
Replay [5]	ASVspoof 2017
SS, VC, and Replay [7]	ASVspoof 2019
Replay [8]	ReMASC

2.3.1 ASVspoof 2015 Challenge Database

The ASVspoof 2015 Challenge database was the first major release for spoofing, and countermeasures research in the context of ASV [38]. The database consists of natural and spoofed speech, which is generated via speech synthesis and voice conversion, for logical access (LA) attacks. There are no remarkable channel or background noise effects. The database is divided into three subsets, namely, training, development, and evaluation. The training set is given to prepare the corresponding genuine and spoof models. These models are used to classify the genuine and spoof speech signals of development and evaluation sets, respectively. We used the development set to train the fusion parameter for that particular database. For example, the development set is used to obtain the fusion parameter and the same parameters are used for evaluation set. The evaluation subset consists of *known* and *unknown* attacks. They include the same 5 algorithms used to generate the development dataset and hence, called as *known (S1-S5)* attacks. In addition, other spoofing algorithms are included in *unknown (S6-S10)*, attacks which were used directly in the test data. The number of speakers in the database is reported in Table 2.3. The detailed description of the database can be found in [2, 38, 85].

Table 2.3: A Summary of ASVspoof 2015 Challenge Database. After [2]

Subset	# Speakers		# Utterances	
	Male	Female	Genuine	Spoof
Training	10	15	3,750	12,625
Development	15	20	3,497	49,875
Evaluation	20	26	9,404	193,404

2.3.2 AVspoof Database

AVspoof database introduces replay spoofing attacks along with synthetic speech, and voice conversion spoofing attacks. It was designed to simulate the attacks via logical and physical access. This database was used in the BTAS 2016 Challenge [3, 100]. The statistics of the database are summarized in Table 2.4. This database reports a comprehensive variety of presentation attacks including attacks, when a genuine data is played back to an ASV system using laptop speakers, high quality multimedia speakers, and two mobile phones. Synthetic speech attacks, such as speech synthesis, and voice conversion replayed with laptop speakers, are also included [100]. The 'unknown' attacks were introduced in the test set in order to make the competition more challenging [100]. The organizers of the challenge provided a baseline system, which is based on the open source *Bob toolbox* [100]. The baseline system consists of simple spectrogram-based ratio as features and logistic regression as a pattern classifier [100].

Table 2.4: A Summary of AVspoof Database. After [3]

Subset	# Utterances		
	Genuine	PA Attacks	LA Attacks
Training	4,973	38,580	17,890
Development	4,995	38,580	17,890
Evaluation	5,576	43,320	20,060

PA: Physical Access, LA: Logical Access

2.3.3 RedDots Replayed Database

RedDots database is developed originally for text-dependent ASV research that was re-developed from the replay attacks [4]. This database is derived from the original RedDots database under various recording and playback conditions. The original RedDots corpus serves as the genuine speech and its replayed version serves as a spoofed data. The spoofed data was recorded in different environments in the European Union Horizon 2020-funded OCTAVE project [104]. The efforts were made to align with the text-dependent ASV and thus, is well positioned for the assessments of replay spoofing countermeasures. The statistics of the RedDots replayed database are reported in Table 2.5.

Table 2.5: A Summary of the RedDots Replayed Database. After [4]

Subset	# Speakers	# Utterances	
	Male	Genuine	Spoof
Training	10	1,508	9,232
Evaluation	25	2,346	16,067

2.3.4 ASVspoof 2017 Challenge Database

The ASVspoof 2017 Challenge database was built on the RedDots corpus [105], and its replayed version [4], which is therefore a replay database, and the database is text-dependent. The number of speakers in training, development, and evaluation subsets with corresponding number of genuine and spoofed utterances are summarized in Table 2.6. The detailed description of the database can be found in [5, 6].

Table 2.6: A Summary of ASVspoof 2017 Challenge Version 2.0. After [5, 6]

Subset	# Speakers	# Utterances	
		Genuine	Spoofed
Training	10	1,507	1,507
Development	8	760	950
Evaluation	24	1,298	12,008

The version 2.0 database presents in depth analysis of the replay detection performance along with description of playback and recording devices. Furthermore, ASVspoof 2017 challenge version 2.0 database was released to correct data anomalies that were detected in the post evaluation of version 1.0 database [6].

Along with the corrected data, more detailed description of recordings, and playback devices as well as acoustic environments was also reported.

2.3.5 ASVspoof 2019 Challenge

The ASVspoof 2019 challenge is an extension of the previously held two challenges (in 2015 and 2017), which focuses on countermeasures for all the three major attack types, namely, synthetic speech, voice conversion, and (synthetic) replay. In particular, there are two sub-challenge, namely, Logical Access (LA), and Physical Access (PA). The statistics of the database is summarized in Table 2.7 [7]. The training dataset includes genuine and spoofed speech from the 20 speakers (8 male and 12 female). The spoof speech signals are generated using one of the two voice conversion and four speech synthesis algorithms. The data conditions for earlier ASVspoof 2017 challenge were created in an uncontrolled setup and hence, this condition made the results challenging to analyze the signal due to varying additive and convolutive noise. This uncontrolled condition was taken care in this challenge by creating a simulated and controlled acoustic environment conditions. Unlike previous challenge editions, ASVspoof 2019 adopts a recently-proposed Tandem Detection Cost Function (t-DCF) as the primary performance metric along with % EER [30, 106] .

Table 2.7: The Summary of ASVspoof 2019 Challenge Database. After [7]

	# Speakers		# Utterances			
Subset	Male	Female	Logical Access (LA)		Physical Access (PA)	
			Natural	Spoof	Natural	Spoof
Training	8	12	2,580	22,800	5,400	48,600
Development	8	12	2,548	22,296	5,400	24,300
Evaluation	–	–	71,747		137,457	

2.3.6 ReMASC Database

The ReMASC (Realistic Replay Attack Microphone Array Speech Corpus) is the first publicly available database that is designed specifically for the protection of Voice Assistants (VAs) (also known as Intelligent Personal Assistants (IPA)) against various replay attacks in various acoustical conditions and environments [8]. The ASVspoof 2019 challenge consists of simulated data for clear theoretical analysis of audio spoofing attacks in physical environments, however, it brings a simulation-to-reality gap [8]. Recent increase for the use of VAs depends on voice input as the primary user-machine interaction modality, such as IPA (e.g., Amazon Echo, Samsung Bixby, Microsoft Cortona, and Google Home) allow users to control their smarthome appliances, and complete many other tasks with ease. The VAs also began to be used in vehicles to allow drivers to control their cars' navigation systems, and other vehicle services. The number of speakers and the acoustical environmental conditions are summarized in Table 2.8. Recently, VAs or IPAs are used for conversational in-vehicle systems.

Table 2.8: Data Volume of the ReMASC Corpus (* Indicates Incomplete Data Due to Recording Device Crashes). After [8]

Acoustic Environment	# Subjects	# Genuine	# Replayed
Outdoor	12	960	6,900
Indoor 1	23	2,760*	23,104
Indoor 2	10	1,600	7,824
Vehicle	10	3,920	7,644
Total	55	9,240	45,472

2.4 Countermeasures for SS and VC Spoofing Attacks

We now give an overview of system construction for anti-spoofing against synthetic speech that includes synthesized, and converted voices. The ASVspoof 2015 challenge provided a common platform to study the effectiveness of countermeasures. Similar to other pattern classification system, a traditional spoof detection system consists of two parts, namely, feature extraction, and pattern classifier (as shown in Fig. 2.5). We will discuss the traditional approach and the end-to-end approach in more detail in this Section.

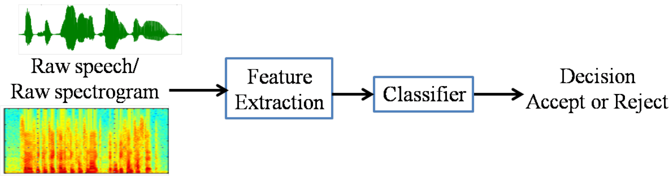


Figure 2.5: Framework for Spoof Speech Detection (SSD). After [1].

2.4.1 Handcrafted Features

There have been several earlier studies on extracting features that reflect the artifacts in the synthetic speech. For example, one study considers that the pitch or fundamental frequency (F_0) pattern of synthetic speech is more rigid than that of natural speech in [107], temporal structure of synthetic speech is different from that of natural speech [108], and synthetic speech contains phase distortions [109]. However, these features were observed on adhoc databases and moreover, they were not evaluated using a common performance evaluation metric. Hence, there was a need to develop a shared task for synthetic speech detection, that motivated the ASVspoof 2015 Challenge [2,38].

In ASVspoof 2015 Challenge, it was observed that the efforts on better features were more effective than the complex classifiers [13]. The Constant-Q Cepstral Coefficients (CQCC) [110], and Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCC-IF) [111] offer state-of-the-art performance on ASVspoof 2015 database. The CFCC-IF feature extraction, that is developed by Speech Research Lab at DA-IICT Gandhinagar represents relatively lowest equal error rate (in terms of ranking) in ASVspoof 2015. The CQCC features are extracted with the constant-Q transform (CQT), a perceptually-inspired alternative to Fourier-based approaches for time-frequency analysis. The CQCC feature set has been reported to perform well on three different databases (i.e., ASVspoof 2015 Challenge, AVspoof, and RedDots replayed database), and it delivered the state-of-the-art performance in each case [112].

Other effective features include high-dimensional magnitude spectrum-based features, and phase-based features as reported in a comparative study [113]. The magnitude spectrum-based features include Log-Magnitude Spectrum (LMS), and Residual Log-Magnitude Spectrum (RLMS); the phase-based features include Group Delay Function (GDF), Modified Group Delay Function (MGDF), Baseband Phase Difference (BPD), Pitch Synchronous Phase (PSP), and Instantaneous Frequency Derivative (IFD).

The features extracted using subband processing were also explored, such as Linear Frequency Cepstral Coefficients (LFCC) [114], Energy Separation Algorithm-Instantaneous Frequency Cepstral Coefficients (ESA-IFCC) [13], and Constant-Q Statistics-plus-Principal Information Coefficient (CQSPIC) [115]. The basic motivation behind subband processing is that artifacts of synthetic speech manifest differently in different subbands. Temporal features, such as instantaneous frequency, and envelope are sensitive to those artifacts. Another technique for subband processing is to perform a two-level scattering decomposition through a wavelet filterbank to derive a scalogram [116].

2.4.2 Representation Learning (RL) Literature

The RL approaches work either in form of feature learning or as a pattern classifier. With feature learning, it was observed that the use of DNN for RL followed by GMM or SVM classifier was more successful than using DNN as a classifier. The hidden layer representation obtained from DNN was used as *features* (called as spoofing vectors or *s*-vectors), and Mahalanobis distance for classification [117]. The CNN and RNN classifiers were explored along with three features, namely, Teager Energy Operator (TEO) Critical Band Autocorrelation Envelope (TEO-CB-Auto-Env), Perceptual Minimum Variance Distortionless Response (PMVDR), and raw spectrograms [118].

In [119], feature learning is followed by LDA, and GMM classifiers. The frame-level and sequence-level features were extracted using DNN and RNN, respectively, resulted in 0 % EER for all the attack types from S1 to S9, and 1.1 % EER on all the averaged conditions [119]. Bottleneck features extracted from the DNN hidden layers were also used with GMM classifier in [120]. In [121], the Convolutional Restricted Boltzmann Machines (ConvRBM) is used for auditory filterbank learning that performed better than the traditionally handcrafted filterbanks. The study reported in [121] shows that ConvRBM learns better low frequency subband filters on ASVspoof 2015 dataset than on TIMIT. Supervised auditory filterbank learning using DNN was also studied in [122]. The first and second-order Long-Term Spectral Statistics (LTSS) were used for SSD task for synthetic speech along with various classifiers with DNN outperforming others [123].

Recently, end-to-end DNN approaches have emerged for various speech and audio processing applications [124], [125]. The goal of the end-to-end DNN is to learn acoustic representation from the raw speech and audio signals as well as perform classification task in a DNN network [126], [127]. SSD task for synthetic speech, Convolutional Neural Network (CNN) was used for feature learning from raw speech signals and binary classification task [128]. Along with CNN layers, Long-Short Term Memory (LSTM) layers were used in an architecture called Convolutional LSTM DNN (CLDNN) trained directly on raw speech signals [129,130]. While CLDNN achieves 0 % EER for the S1-S9 conditions, it has not worked well for S10 set (which is unit selection-based TTS spoof speech). The end-to-end DNN and GANs approach represents a new direction of anti-spoofing study [124], [125].

In ASVspoof 2015 Challenge, the systems in [113] and [131] use DNN as the classifiers. In [132], a DNN classifier with novel human log-likelihoods (HLL) scoring method was proposed that performed better and achieved an average EER of all the attack types to 0.04 %. It was shown in [132] that HLL scoring method is more suitable for the SSD task than the classical LLR scoring method, especially when the spoofed speech is very similar to the human speech [132]. The output softmax layer consists of neurons representing spoofing and human (natural) speech labels. According to the literature, the performance of various anti-spoofing systems on ASVspoof 2015 challenge database is summarized in Table 2.9, with the system of CQCC feature set, and DNN-HLL classifier representing the best performance [132].

Table 2.9: Comparison of Results (in % EER) on ASVspoof 2015 Challenge Database. After [1].

Feature Set	Classifier	Dev	Eval
CQCC [112]	GMM	0.00	0.26
CFCCIF [111]	GMM	2.29	1.21
LFCC [114]	GMM	0.66	0.89
RFCC [114]	GMM	075	1.02
MFCC [114]	GMM	1.09	3.0
SCFC [114]	GMM	0.25	4.45
SCMC [114]	GMM	0.95	0.94
LPCC [114]	GMM	0.68	1.21
IMFCC [114]	GMM	0.48	1.00
RPS [114]	GMM	0.37	5.30
SCC [116]	GMM	-	0.18
DMCC-BNF [120]	GMM	-	2.15
ESA-IFCC [13]	GMM	1.89	6.79
ConvRBM-CC [121]	GMM	2.53	4.47
DNN-IGFCC [122]	GMM	0.12	0.56
LF RPS [131]	SVM	1.34	6.11
DNN, RNN features [119]	LDA, GMM	-	1.1
LF Spectrum [131]	DNN	0.03	4.38
TEO [118]	DNN	2.31	-
PMVDR [118]	DNN	1.44	-
LTSS [123]	DNN	-	0.25
E2E CNN [128]	DNN	-	2.89
LTSS and E2E CNN [128]	DNN		0.157
E2E CLDNN [130]	DNN	-	4.56
CQCC [132]	DNN-HLL	-	**0.04**
Spectrogram [118]	CNN	0.36	3.07
Spectrogram [118]	RNN	1.04	2.46
Spectrogram [118]	CNN+RNN	0.42	1.86

2.5 Countermeasures for Replay Spoofing Attacks

We now present an overview of literature for anti-spoofing (or countermeasures) against replay attacks.

2.5.1 Handcrafted Features

The use of high-fidelity recording devices represents a serious threat and hence, countermeasures were proposed to guard against such attacks. The spectral peak mapping method was proposed as a countermeasure to detect the replay attack on a remote telephone interaction [133]. Replay attacks with far-field recordings were addressed in [50].

The ASVspoof 2017 Challenge paid a special attention to replay speech detection. The baseline system with CQCC features and GMM classifier was provided by the organizers as it performs well in the earlier evaluation [5]. The acoustic features, such as Rectangular Filter Cepstral Coefficients (RFCC), Subband Spectral Centroid Magnitude Coefficients (SCMC), Subband Spectral Centroid Frequency Coefficients (SCFC), and Subband Spectral Flux Coefficients (SSFC) were studied for replay SSD task. It is found that the SCMC followed by feature normalization method outperforms other acoustic features [134]. With the analysis on Inverse Mel Frequency Cepstral Coefficients (IMFCC), Linear Prediction Cepstral Coefficients (LPCC), and LP Residual features, it is found that high frequency regions have more discriminative information than the other frequency regions [135]. The effect of mean and variance normalization of CQCC feature set with Support Vector Machines (SVM) classifier was studied in [6,136]. One of the approaches used Single Frequency Filtering (SFF), and found the importance of high resolution temporal features [137].

The short-time AM-FM features set obtained using Energy Separation Algorithm (ESA) were studied in [14,28]. The features were also developed with subband filter analysis using CFCC-IF [28], IFCC [138], Empirical Mode Decomposition Cepstral Coefficients (EMDCC) [139], transmission line cochlear model [140], auditory inspired spatial differentiation filterbank [141], and ESA-IF-based feature estimation using cochlear filter in [142]. Excitation source-based features were studied in [143], wavelet-based features in [144], and phase-based features in [145]. The concept of feature switching at the decision-level, along with information from the non-voiced segments was studied in [146].

The study in [147] shows that some phonemes carry more replay artifacts than the others and thus, judicious use of phoneme-specific models can improve replay detection. The analysis of full-frequency bands with F-ratio, and multi-channel feature extraction using attention-based adaptive filters (AAF) is studied in [148]. The analysis of replay speech signal using *reverberation* concept, and Teager energy profile is studied in [11].

2.5.2 Representation Learning (RL) Approaches

The three key observations from ASVspoof 2017 Challenge are the use of spectral information in the higher frequency regions, feature normalization, and representation learning approach. It was shown that many representation learning-based approaches did well in the ASVspoof 2017 Challenge.

First, we describe the RL approaches used in ASVspoof 2017 Challenge. End-to-end replay spoofing detection was proposed using deep residual network (ResNet), and raw spectrograms of speech signals [149]. It was also shown that the data augmentation in DNN significantly improves the performance [149]. In one of the approaches, DNN was trained to discriminate between the various channel conditions available in the ASVSpoof 2017 challenge database, namely, recording, playback, and session conditions [150]. In [150], the DNN features were learned from CQCC and HFCC features followed by an SVM classifier. The model fusion strategies using ResNet, GMM, and DNN were also explored and found to perform better compared to the individual systems [151]. In particular, the ASVspoof 2017 Challenge winner system used CNN and RNN for representation learning from STFT spectrograms followed by a GMM classifier [152].

The use of ConvRBM to learn auditory filterbank followed by the AM-FM demodulation using ESA for the replay SSD task was studied in [153]. The ConvRBM learns subband filters that represent high frequency information in a much better way when used with pre-emphasized speech signals. Combining representation learning and signal processing techniques gives significant improvement of 0.82 % and 8.89 % EER on the development and evaluation sets, respectively. A novel algorithm called NeuroEvolution of Augmenting Topologies (NEAT) was used in an end-to-end anti-spoofing network [154]. The NEAT framework also introduces a new fitness function for DNN that results in better generalization than the baseline system, and improves the relative performance by 22 % on the ASVspoof 2017 database [154]. A novel visual attention mechanism is employed in deep ResNet architecture using the group delay function features (GD spectrum) that resulted in 0 % EER on both the development and evaluation sets, respectively [155]. In [156] attention-based filtering is used that enhances the feature representation in both time and frequency-domains and used ResNet-based classifier. Class Activation Maps (CAM) using Global Average Pooling (GAP) utilizes the implicit attention mechanism present in CNN. Hence, representation learning approaches are very promising directions for the replay SSD compared to the synthetic SSD task. According to the literature, the performance of various replay anti-spoofing systems on ASVspoof 2017 challenge dataset is summarized in Table 2.10.

2.6 Limitations and Challenges

In this Section, we discuss limitations and technological challenges that are worthy of further inquiry and possible future directions.

1. **Why Replay Detection is Challenging ?**
   Replayed speech is expressed via a convolutional model under assumptions of LTI systems. In order to detect replay speech, we need to capture the impulse response of the intermediate device, acoustic environment. Thus, replay speech detection is a *blind* deconvolution problem and hence, getting exact *deconvolution* of $h(n)$ from $r(n)$ is still a challenge in signal processing .

2. **Diversity of Spoofing Attacks**: The ASVspoof 2015 challenge database was designed only voice conversion (VC), and synthetic speech (SS) spoofing algorithms. This database consists of variation of seven voice conversion spoofing techniques and only three synthetic speech generation techniques. It is noted that ASVspoof 2017 challenge database focuses only on replay spoof. While ASVspoof 2019 database includes synthetic speech (SS), voice conversion (VC), and replay spoofing techniques, the spoofing voice database is not developed using the latest neural voice generation techniques.

3. **Number of Speakers**: Different spoofing databases consists of a different number of speakers either male, female or both. The number of speakers present in ASVspoof 2015 challenge database consists of a large number of speakers (male and female), whereas for ASVspoof 2017 challenge database, only male speakers where considered. The studies have reported that when the number of speakers used during training is increased, it improves the performance in % EER. However, performance changes with the spoofing attacks and also with the features that are used during training. Hence, the performance measures should also justify the independence towards the number of speakers, and the voice of the speaker under consideration.

Table 2.10: Comparison of Results (in % EER) on ASVspoof 2017 Challenge Database. After [1]

Feature Set	Classifier	Dev	Eval
CQCC (BL) [5]	GMM	10.35	28.48
CFCC-IF [28]	GMM	6.80	34.49
EMDCC [139]	GMM	28.48	28.06
AFCC [148]	-	4.01	27.80
SCFC [134]	GMM	24.51	24.83
HFCC [150]	GMM	5.9	23.90
PNCC [157]	GMM	20.78	23.74
CQCC [149]	ResNet	6.32	23.14
LFRCC [143]	GMM	8.38	22.28
SSFC [134]	GMM	12.81	22.38
SCC [144]	GMM	3.16	19.79
CQCC [151]	DNN	5.18	19.41
DLFS [146]	GMM	6.68	19.16
CQCC [151]	ResNet	5.05	18.79
EOC_m [154]	-	-	18.2
LPCC [158]	GMM	-	17.0
LFCC [134]	GMM	10.31	16.54
VESA-IFCC [28]	GMM	4.61	14.06
ESA-IFCC [27]	GMM	4.12	12.79
ARP [148]	-	9.11	12.65
LPCC [152]	SVM *i-vector*	9.80	12.54
VESA-IACC [14]	GMM	6.12	11.94
TECC [11]	GMM	9.55	11.73
AWFCC [29]	GMM	6.37	11.72
SCMC [134]	GMM	9.32	11.49
CF [141]	GMM	-	10.84
ESA-IFCC [58]	GMM+CNN	1.90	10.42
PPWS [147]	GMM	-	10.70
PPRFWS_LR [147]	GMM	-	9.28
AF-DRN [156]	ResNet	6.55	8.99
ConvRBM-CC [153]	GMM	0.82	8.89
TLC_AM [140]	GMM	-	8.68
FFT Features [152]	LCNN	4.53	7.37
GD Spectrum [155]	ResNet	**0.0**	**0.0**

BL: Baseline; -:Information not found

4. **Signal Degradation Conditions**: Current publicly available spoofing databases are developed in clean conditions. However, the ASVspoof 2017 replay

database was recorded under various acoustic environmental conditions. For ASVspoof 2015 challenge database, the noisy database was developed by adding various noises at different Signal-to-Noise Ratio (SNR) levels. Further investigations are required as to how the diversity of different noise types affects the SSD performance. In addition, the study is required to observe the effect on SSD, when the additive noise is added manually, and when the noise is added naturally via the acoustic environment. Study for different acoustical background, microphone. is reported in [6]. Hence, the countermeasures must be developed that it should be robust to signal degradation conditions as well.

5. **Robustness in ASV Implies Vulnerability**:
   In practice, we would like an ASV system to be robust against variations, such as microphone and transmission channel, intersession, acoustic noise, speaker aging. A robust ASV system may become vulnerable to various spoofing attacks as it tries to nullify these effects, and normalize the spoofing speech towards the natural speech. Thus, robustness and anti-spoofing security should be addressed separately. It is worth to study how features, classifiers, and systems are designed to be both robust and secure.

6. **Exploiting of Specific Frequency Region: Why ?**
   In practical scenarios, a replay might be done by and enlarge in air medium that contains the air particles having mass and springiness and thus, a slug of air will be responsive to a particular frequency band, which will emphasize onto the spectrum of the replayed speech. The investigation for replay detection has shown the significance of selecting a particular frequency regions [135,150].

7. **Lack of Exploiting Excitation Source Information**:
   Less amount of work is done in using excitation source assuming that the Glottal Closure Instants (GCI) are having sharp impulse-like nature for voiced speech. The spectrum of the glottal source (i.e., Glottal Flow Waveform (GFW)) for voiced speech is expected to have *harmonic* structure in the frequency-domain. Thus, any deviation from the degradation in the harmonic structure could capture the signature of spoofed speech [143]. To the best of author's knowledge, there is no study reported in analyzing this particular aspect. We believe several excitation source information, such as Linear Prediction (LP) Residual [159], Teager Energy Operator (TEO) profile, and its Variable length version, i.e., Variable length TEO (VTEO) profile. could be

explored in the framework of recent study reported in [143].

8. **Exploring Phase-based Features**:
   It is important to note that the phase-based features (either time-domain analytic or frequency-domain) could capture different kind of information in spoofed speech depending upon the type of spoof. For example, in USS system, when the speech sound units are picked up by optimizing the target cost, in the synthesized voice, might have *linear phase* mismatches (if these units are recorded in different sessions) [160]. On the other hand, for replayed speech, the impulse response of the acoustic environment (say room) gets convolved with the natural speech. The impulse response of an acoustic system (in this case room) is infinite in duration, i.e., it is Infinite Impulse Response (IIR) in nature (due to infinite transmissions and reflections in the given acoustic room). Thus, the nonlinear phase in frequency-domain of this acoustic system is added to the phase of natural speech. In addition, corresponding effects of this nonlinear phase could be observed in temporal-domain, such as non-integer delay in frequency components. There have been many studies in phase features in synthetic speech detection. Phase study remains a research topic that is worth for further investigations.

To alleviate the problems discussed above, this thesis proposes novel front-end feature sets, which is shown to perform better than the baseline systems for the SSD task [11, 27]. Later, the proposed feature sets are applied on various audio classification tasks, such as Acoustic Scene Classification (ASC) [23], countermeasures for replay SSD for Voice Assistants (VAs) or Intelligent Personal Assistants (IPA) [20], Whisper Speech Detection (WSD) [22], and Automatic Speech Recognition (ASR) tasks for near-field *vs.* far-field scenarios [19]. The novelty of the proposed feature sets lies in the technique used for speech demodulation, i.e., ESA and TEO approaches. The feature extraction process is done in time-domain. In addition, the ESA approach do not require the complex task of phase unwrapping. The slow and fast-varying temporal modulations are accurately estimated using ESA technique. The Teager energy profiles of the narrowband filtered signals shows the nonlinearity around the GCI locations because of the speech production mechanism, and shows the difference between the natural, and its corresponding spoof speech signal. Earlier the TEO approach was applied for speech and speaker recognition, emotion recognition, and speech enhancement tasks [161–164]. In this thesis, we explored different applications using TEO and ESA.

2.7 Chapter Summary

This Chapter presented the motivation and literature towards choosing replay, SS, and VC spoof detection for analysis of the security threat to the ASV systems. In addition, the various types of spoof speech signals were presented briefly. The discussion on different approaches proposed for detection of spoof speech signal along with the corpus used in this field are also discussed. The various issues with current approaches are discussed that needs to be addressed in the near future. In the next Chapter, the spoofing techniques and spoof detection system along with the various databases used in this thesis, classification system, and the performance measures for the evaluation of countermeasures in the SSD system are discussed.

CHAPTER 3

Teager Energy Operator (TEO)

3.1 Introduction

In this Chapter, we introduce our proposed feature set based on the Teager Energy Operator (TEO), namely, Teager Energy Cepstral Coefficients (TECC). The concept of the proposed approach is discussed in Section 3.2 and the basic concepts of nonlinearity in the natural speech production estimated by the Teager energy profiles is discussed in Section 3.3. In addition, the feature extraction process is discussed in Section 3.4. Furthermore, the analysis of Teager energy profiles of spoof speech signals is discussed in Section 3.5. We studied modulations of energy estimated via TEO profile to emphasize the impulse that arrives because of echo/reverberation in Section 3.6. An application to the SSD task is presented in Section 3.7-3.8 for various databases. Finally, in Section 3.9, we summarized the Chapter.

3.2 Basics of TEO

According to Newton's second law of motion, for an oscillator with mass, m, and spring constant, k, its displacement $x(t)$ is governed by the motion equation, for which the general solution is a cosine $x(t) = Acos(\omega t + \phi)$, where A is the amplitude of the oscillation, t is continuous time-domain with $\omega = \sqrt{k/m}$ is frequency of oscillations (rad/sec), and ϕ is the initial phase (rad), we have following differential equation [57]:

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0. \tag{3.1}$$

The instantaneous energy E, of this undamped oscillator is constant and equal to the sum of its kinetic and potential energy, i.e.,

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2 \Rightarrow E = \frac{1}{2}m\omega^2 A^2, \tag{3.2}$$

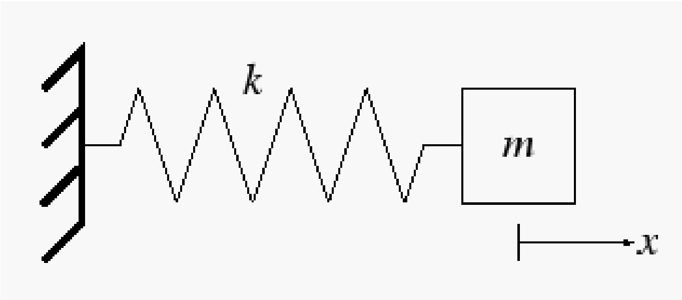


Figure 3.1: Schematic of Ideal Simple Harmonic Motion (SHM).

where $\omega = \frac{d}{dt}\phi(t)$. An algorithm derived by Teager uses a nonlinear energy tracking operator [52]. The n^{th} sample of a discrete-time monocomponent signal can be expressed as:

$$x(n) = Acos(\omega n + \phi), \tag{3.3}$$

from Eq. (3.3) we can write:

$$x(n+1) = Acos(\omega(n+1) + \phi), \tag{3.4}$$

$$x(n-1) = Acos(\omega(n-1) + \phi). \tag{3.5}$$

When we multiply above equations, we obtain,

$$x(n+1)x(n-1) = A^2cos(\omega(n+1) + \phi)cos(\omega(n-1) + \phi), \tag{3.6}$$

$$x(n+i)x(n-i) = [Acos(\omega n + \phi)]^2 - A^2sin^2\omega. \tag{3.7}$$

Using Eq. (3.3) in Eq. (3.7),

$$A^2sin^2\omega = x^2(n) - x(n-1)x(n+1). \tag{3.8}$$

For small values of ω, $sin\omega \approx \omega$, and hence, Teager Energy Operator (TEO), $\Psi_d\{\cdot\}$, is defined as [165]:

$$E_n = A^2\omega^2 \approx x^2(n) - x(n-1)x(n+1) = \psi\{x[n]\}, \tag{3.9}$$

where E_n gives the running estimate of signal's energy, A is the amplitude, and ω is frequency (in radians). The speech signal is the combination of several monocomponent signals [51]. Considering the speech signal, the TEO cannot be applied directly on the speech signal as it is the summation of multicomponent signals. In order to obtain a narrowband signal, the speech signal is passed through a bandpass filtered signal in order to obtain N number of subband signals [53]. The subband signal at center frequency is obtained from $r_i(t) \approx s(t) * g_i(t)$, $s(t)$ is the speech signal, and $g_i(t)$ is the impulse response of the i^{th} Gabor filter. The impulse response of the Gabor filter is given as [52]:

$$g(t) = exp(-b^2t^2)cos(\omega_c t). \tag{3.10}$$

Here, Gabor filter acts as a bandpass filter with center frequency, ω_c. In this Section, Gabor filter is used placing the center frequency to linear scale, so that the Gabor filters are equally distributed into the entire frequency range [13,166,167].

3.3 Analysis of TEO Profile

In our earlier study [11], we tried to link the concept of reverberation with replay SSD task, as the replay signal are recorded and played back, where the reverberation exist. In Figure 3.2, the synthetic sinusoidal signals (Panel I) are shown along with their corresponding TEO profiles (Panel II). Figure 3.2(a) show the damped sinusoidal signal with the peak amplitude of impulse is equal and Figure 3.2(b) show the damped sinusoidal signal with decrease in amplitudes of the impulse. Whereas, Figure 3.2(c) show the variations in the amplitude of the damped sinusoidal signal. It can be observed from their corresponding TEO profiles in Panel II that for each case the TEO profile show impulse-like energies. In particular, if the peak amplitude of the signal is constant, the TEO profiles are also constant in terms of its amplitude, and if the amplitude of signal varies (as in case on Panel I (b and c)), the corresponding TEO profiles also varies (highlighted by the box and oval shapes).

The TEO profiles show high energy pulses around the Glottal Closure Instant (GCI), because of impulse-like excitation to vocal tract system and this sudden glottal closure produces high energy and thus, TEO produces high energy around these regions due to the high time resolution property of TEO to estimate the signal's energy [168]. Along with high Teager energy pulses, the *bumps* are observed indicating contributions for nonlinear effects during the speech production process [168]. This nonlinear effect is observed for real speech signal as shown in Figure 3.3, in particular, for natural (Figure 3.3(a)) and its corresponding replay speech signal (Figure 3.3(b)). When compared to the synthetic signal as shown in Figure 3.2 the nonlinearities around the GCI locations are missing and hence, the natural speech confirms the capability of TEO to represent characteristics of airflow pattern during natural speech production.

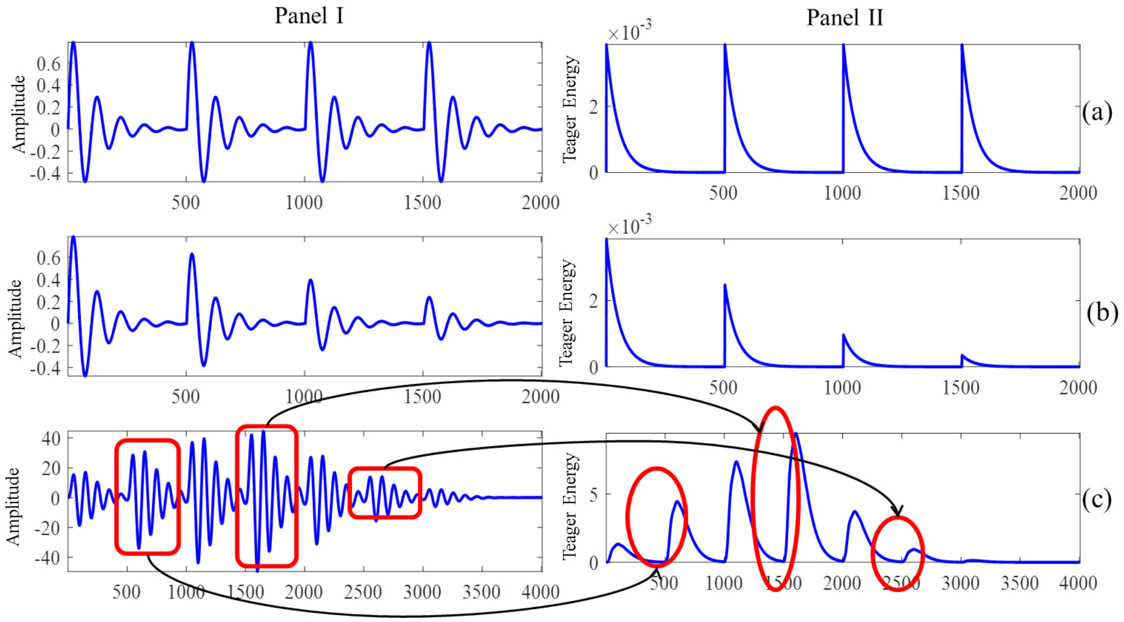


Figure 3.2: Panel I: Synthetic signals with (a) same, (b) decreasing, and (c) varying sinusoidal signals along with their corresponding Teager energy profiles in Panel II. After [16].

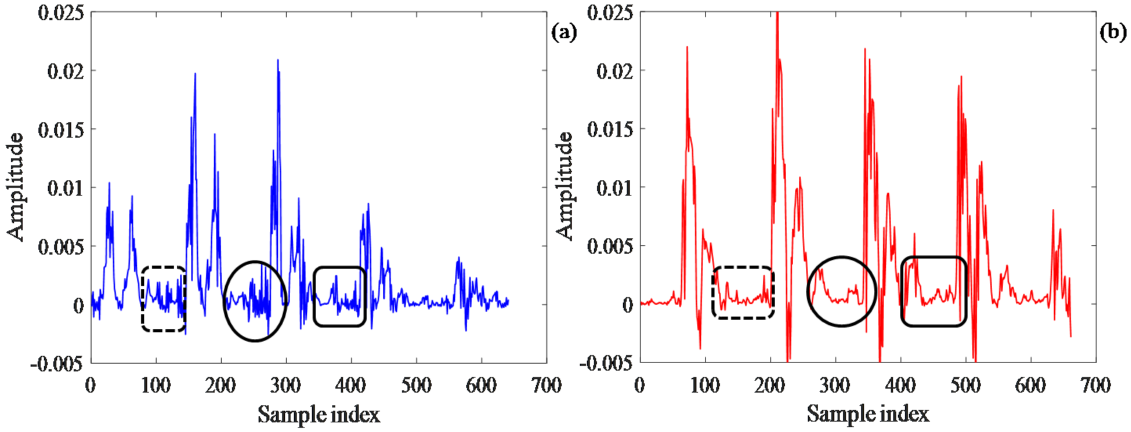


Figure 3.3: Teager energy profiles for (a) natural, and (b) replay speech segment. Highlighted regions shows the contribution of nonlinear effects during speech production process which is not observed for synthetic case. After [16].

3.4 Feature Extraction Process

The functional block diagram of the proposed TECC feature extraction is shown in Figure 3.4. Here, the input speech signal is passed through a pre-emphasis filter having a system function, $H(z) = 1 - az^{-1}$, with a typical value of $a = 0.97$ [169], to emphasize high frequency regions [135]. This pre-emphasized speech signal is then passed through a Gabor filterbank in order to obtain narrowband filtered signals. The Gabor filter is compact, smooth, and also has *optimal* joint time-frequency resolution, and thus, distortions and noise present in distinct locations, time or frequency, do not interfere with the filter responses [170]. The optimal criteria here is to be able to achieve minimum time-bandwidth product that is dictated by Heisenberg's uncertainty principle in signal processing framework [170]. In particular, following statement. The temporal variance and the frequency variance of a signal, $f(t) \in L^2(R)$ (i.e., Hilbert space of square integrable functions) satisfy,

$$\sigma_t^2.\sigma_\omega^2 \geq 1/4. \tag{3.11}$$

This inequality becomes equality if and only if $f(t)$ is a Gaussian, where $\sigma_t^2.\sigma_\omega^2$ is called as time-bandwidth product (which is also area of Heisenberg box). Studies in [27] found that the linearly-spaced center frequencies have high resolution in both the lower and higher frequency regions that make more reliable to estimate the spectral information. Hence, the narrowband filtered signals are obtained at center frequency, which are linearly-spaced between f_{min}=10 Hz, and f_{max}=8000 Hz. The Gabor filter has the linear phase response characteristics and hence, it maintains the same pattern (shape) of the filtered speech signal (within the passband of filter) with a delay in time which is equal to group delay function (in seconds) of the filter [171]. The center frequency for ERB and Mel have number of cut-off frequencies in the lower frequency regions. Motivated by the studies of auditory perception mechanism for humans, the center frequencies of ERB and Mel scales have narrow and wider bandwidth in the lower and higher frequency regions, respectively [172, 173]. In case of linear scale, all the subband filters have almost *equal* bandwidth and hence, have high resolution in the lower and higher frequency regions that make more reliable to estimate the spectral information. The filtered subband signals obtained from the linearly-spaced Gabor filterbank are applied to the TEO block, and estimate the instantaneous energy of each subband filtered speech signal. Furthermore, these Teager energy profiles are passed through the frame-blocking, and averaged with a short window of 20 ms and with a window shift of 10 ms followed by logarithm operation to compress the data. To obtain a low-dimensional representation that has compact energy, Discrete Cosine Transform (DCT) is applied along with Cepstral Mean Normalization (CMN) (also known as Cepstral Mean Subtraction (CMS)) to reduce the channel mismatch/distortion conditions [174]. Finally, retained few DCT, i.e., Teager Energy Cepstral Coefficients (TECC) appended along with their Δ and $\Delta\Delta$ features to obtain higher-dimensional feature vector.

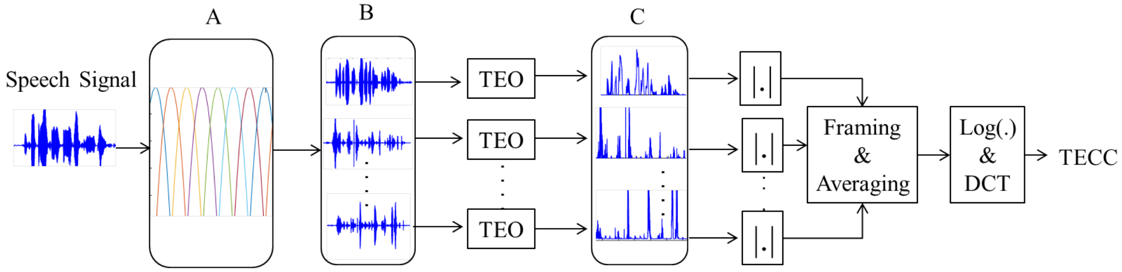


Figure 3.4: Block Diagram of TECC Feature Extraction. A: Gabor Filterbank, B: Narrowband Filtered Signals, and C: Teager Energy Profiles of Each Subband Filtered Signals. After [11].

3.5 Analysis of Spoof Speech Signals

The power spectrum $S_{xx}(f)$ of a time series $x(t)$ describes the distribution of power into frequency components composing that signal. The energy spectral density is suitable for transients (i.e., pulse-like signals) whose energy is concentrated around one time window. For continuous signals over all the time, the power spectral density (PSD) is used this describes how power of a signal or time series is distributed over frequency.

The average power P of a signal $x(t)$ over all time is therefore given by the following time average:

$$P_{av} = \lim_{T\to\infty} \frac{1}{T} \int_0^T |x(t)|^2 \, dt. \tag{3.12}$$

or

$$P_{av} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^T |x(t)|^2 \, dt. \tag{3.13}$$

Fourier transform where the signal is integrated only over a finite interval $[0, T]$ is given as:

$$\hat{x}(\omega) = \frac{1}{\sqrt{T}} \int_0^T x(t) e^{-i\omega t} \, dt. \tag{3.14}$$

This is the amplitude spectral density. Then, the PSD can be defined as:

$$S_{xx}(\omega) = \lim_{T\to\infty} \mathbf{E}\left[|\hat{x}(\omega)|^2 \right]. \tag{3.15}$$

Wiener-Khinchin theorem is stated as the Fourier transform of autocorrelation function. It can also be written with the frequency measured in cycles (rather than radians) per second and denoted by ν.

$$C(\tau) = \int_0^{\infty} 2P(\nu)\cos(2\pi\nu\tau)d\nu = \int_{-\infty}^{+\infty} P(\nu)e^{-2\pi i\nu\tau}d\nu \tag{3.16}$$

We observed the Power Spectral Density (PSD) of natural (blue color), VC (pink color), and SS (red color) signal (from ASVspoof 2015 database) in Fig. 3.5. The PSD shows the stability of energy as a function of frequency, and energy (variations) are strong or they are weak at each frequency [175]. From Fig. 3.5(a), we can see very less difference between natural, and VC PSD plots they approximately overlap on each other, and have minor difference at higher frequency regions. On the other hand, the PSD obtained for natural and SS (as shown in Fig. 3.5(b)) shows very large difference almost for entire frequency regions.

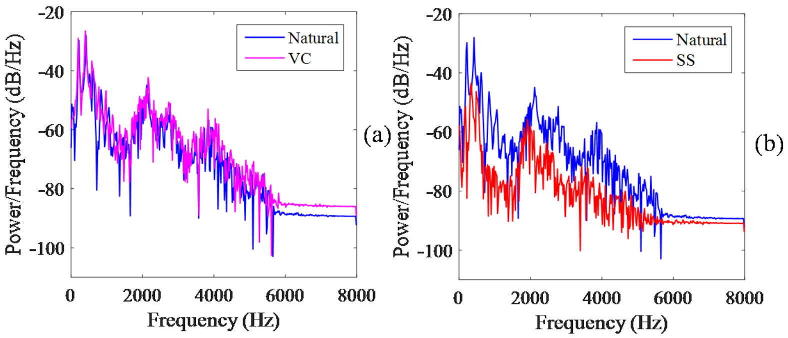


Figure 3.5: Power Spectral Density (PSD) for (a) Natural, *vs.* VC, and (b) for Natural *vs.* SS. After [10].

Furthermore, the Teager energy profiles of the speech segment for natural (Panel I), VC (Panel II), and SS (Panel III) is analyzed as shown in Fig. 3.6. It can be observed that the Teager energy traces obtained for a segment of natural speech signal have more energy, and more bumps are observed corresponding to the Glottal Closure Instant (GCI). Similar observation is found for segment of VC signal. However, the bumps around the GCI locations are very less compared to the Teager energy traces of natural signal. On the other hand, for the segment of SS signal, it can be observed that there are smooth bumps with very less fluctuations (indicating lesser *energy modulations* and non-linearities due to absence of natural speech production activities) in the instantaneous Teager energy traces compared to both natural, and VC bumps. This observation (highlighted with black box and arrows) is the key difference, and it helps to detect the VC and SS spoof signals from the natural speech. In addition, we observed the difference in terms of spectral energies of Teager energy obtained from the output of the Gabor filterbank (as shown in Fig. 3.7). The spectral energy obtained from Teager

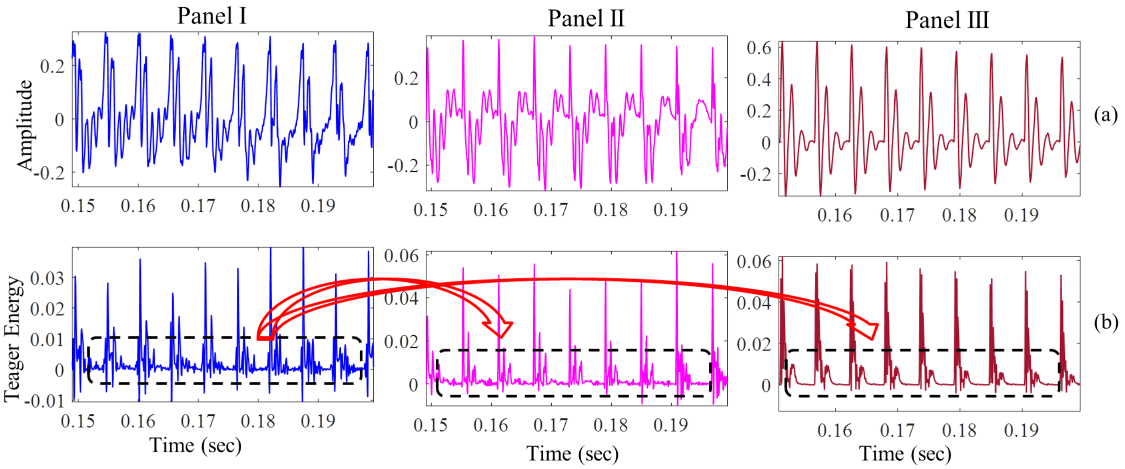


Figure 3.6: (a) Speech Segment of Natural (Panel I), VC (Panel II), and SS (Panel III) Along With Their Corresponding Teager Energy Profiles in (b). Highlighted Regions and Arrows Indicate Change in Teager Energy Bumps (within Two Consecutive GCIs) for All the Cases, In Particular, for Panel III, the Bumps in TEO Profile are Very Smooth. After [10].

energy for the natural speech preserves the formants and harmonics as shown in Fig. 3.7(a). Similar observation for VC signal is found with very less difference in the Teager energy (highlighted by the ovals) as shown in Fig. 3.7(b). The spectral energies obtained from Teager energy for SS signals shows the distorted, and blurred energy compared to the natural speech signal as shown in Fig. 3.7(c). We can see that there is a loss in the energy and harmonics in the higher frequency regions (highlighted with box) in Fig. 3.7(c).

Fig. 3.8 shows the (a) time-domain speech signal, spectral energies obtained from (b) Short-Time Fourier Transform (STFT), and (c) Teager energy-based method for all the speech signals (from BTAS 2016 competition dataset [100]). The Panel I is for natural speech, and corresponding replay signals are shown in Panel II: Played back with Laptop, and Panel III: Played back with Laptop with high quality speaker, Panel IV and Panel V are corresponding synthesized, and voice converted speech signals that are played back with laptop and high quality speaker, respectively. For all the conditions in Fig. 3.8, it can be observed that the spectral energy density obtained from the Teager energy-based approach has high energy across entire frequency regions (because of linearly-spaced Gabor filterbank) as compared to the spectral energy density obtained from the traditional spectrogram. For natural speech signal, the formant frequencies have dark band color showing high energy portions of the speech signal. The shape of the dark bands shows the change from one sound unit to other w.r.t vocal tract shape. When we compare the energies of natural and replay speech signal, the replayed speech

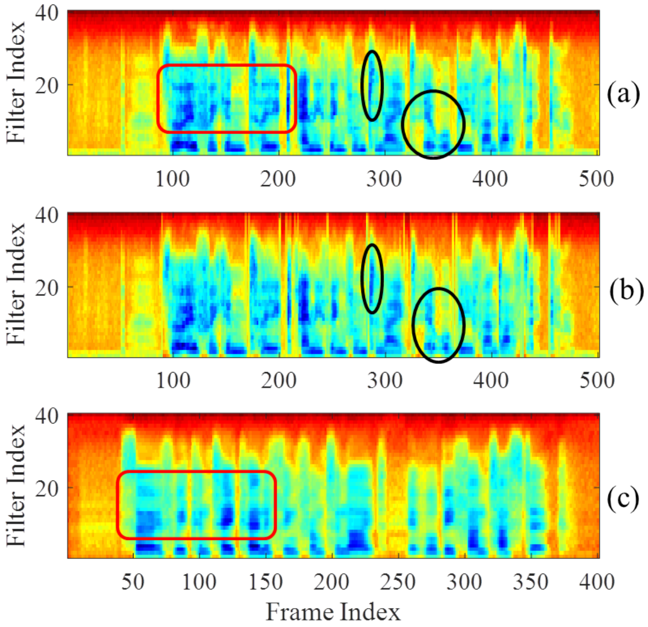


Figure 3.7: Comparison of Teager Energy Features for (a) Natural, (b) VC, and (c) SS Speech Signal from ASVspoof 2015 database. Highlighted Regions via Rectangles and Ellipses Shows the Difference Between the Natural *vs.* VC, and SS. After [10].

obtained with the high quality speaker device (Panel III) has similar pattern of energy, and formant frequency band along with similar time-domain signal pattern. Whereas replay speech with normal quality device (Panel II) has distortions in the energies. For playback speech of machine-generated speech (i.e., Panel IV and Panel V), it can be observed that the spectral cues are not captured with traditional spectrogram Fig. 3.8(b), which is captured with the Teager energy approach and hence, it helps to detect the natural *vs.* spoof speech signals.

The Teager energy profiles for a speech segment of natural (Panel I), replay laptop (Panel II), replay with HQ laptop (Panel III), SS with HQ laptop (Panel IV), and VC with HQ laptop (Panel V) are shown in Fig. 3.9. It can be observed that the Teager energy profiles obtained from various speech signals shows different energy profiles. However, Panel III shows similar pattern of Teager energy traces with natural speech segment, because replay signal is recorded, and replayed with HQ laptop device and hence, it is very similar to the natural counterpart and difficult to detect. It can also be observed from Table 3.5, the HTER for replay is better than the replay with HQ laptop. For Teager energy profiles of SS and VC, we can clearly observe the differences between the natural and replay speech signals. This is also strongly observed from our experimental results showing the lower HTER for SS and VC using HQ laptop as reported in Table 3.5.

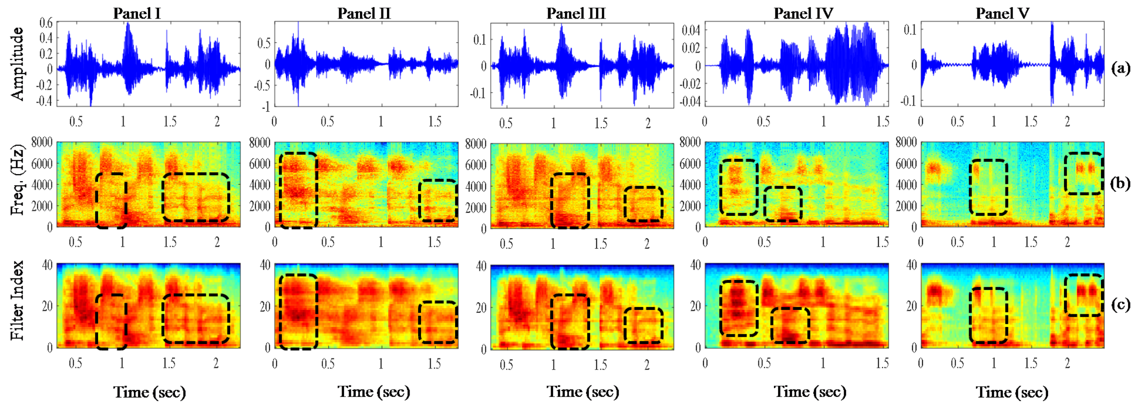


Figure 3.8: (a) Time-domain Speech Signal, Spectral Energy Densities Using (b) STFT Spectrogram, and (c) Teager Energy. Panel I: Natural, Replay Signals Played Back with Panel II: Laptop, and Panel III: Laptop HQ Speaker, Panel IV: Speech Synthesis Physical Access HQ Speaker, Panel V: Voice Conversion Physical Access HQ Speaker. Highlighted Regions Indicates the Discriminative Regions Between the Traditional Spectrum and Teager Energies. After [10].

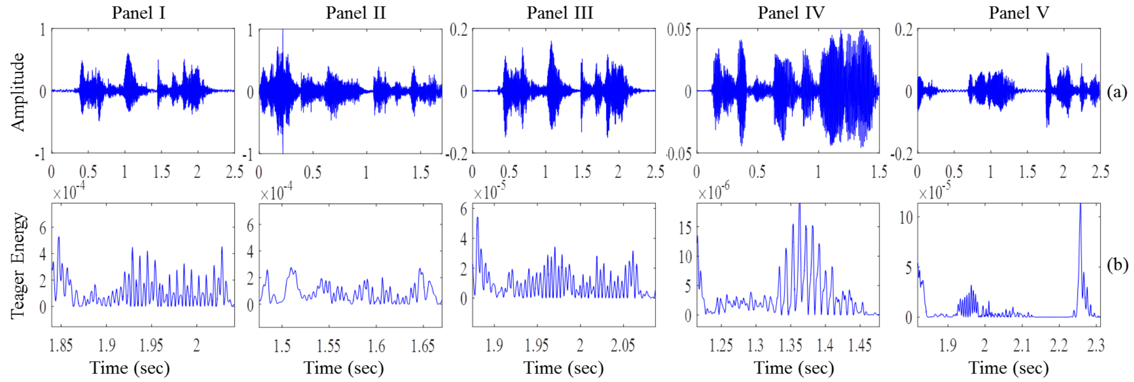


Figure 3.9: (a) Time-Domain Speech Signal, and (b) Teager Energy Profiles of a narrowband signal. Panel I: Natural, Replay Signals Played Back with Panel II: Laptop, and Panel III: Laptop HQ Speaker, Panel IV: Speech Synthesis Physical Access HQ Speaker, Panel V: Voice Conversion Physical Access HQ Speaker. After [10].

3.6 Effect of Reverberation on Replay Speech Signal

3.6.1 Basics of Replay Speech Signal

The task of replay spoof detection is to identify whether a given speech sample is a genuine speech or whether it is recorded version of the genuine speech through an intermediate (recording + playback) devices. The "intermediate device" in this context means that the devices used during the recording and playing it back in order to obtain the replayed speech signal. In particular, during recording different kinds of mic, speaker, tape recorder. are used. The scenario of replay spoof speech detection (SSD) system is shown in Figure 3.10. In particular, the Figure 3.10 shows the process of generation of the replay speech signal in two different acoustic environments, i.e., during recording and playback, different kinds of mic, speaker, tape recorder. are used. The genuine speech signal, $s[n]$, can be modeled

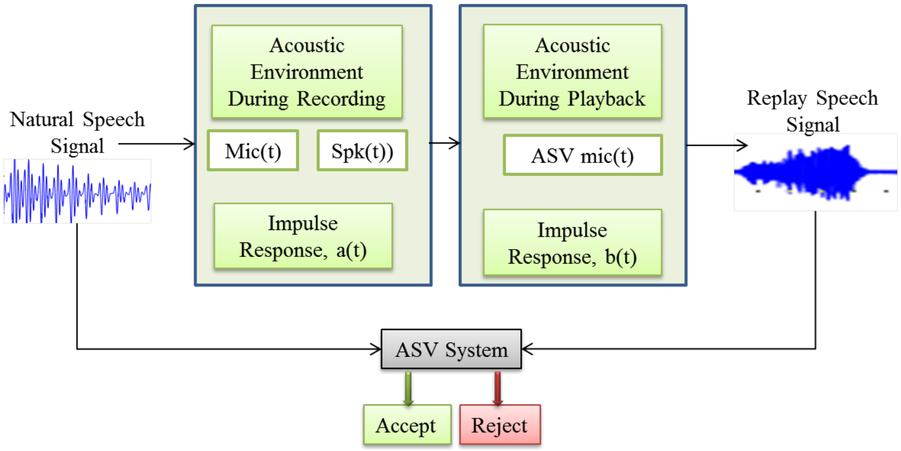


Figure 3.10: Illustration of Replay Spoof Attack Scenario at ASV System. After [9].

as a convolution of glottal airflow, $p[n]$, with the impulse response of vocal tract system, $h[n]$ [175], i.e.,

$$s[n] = p[n] * h[n]. \tag{3.17}$$

It should be noted that as convolution operation requires assumption of linear time-invariant (LTI) system, Eq. (3.17) and subsequent analysis in this work is valid either for a segment (10-30 ms) of speech signal $s[n]$ or the impulse response of vocal tract system is fixed for a speech frame, i.e., $h[n]$ is dependent on the index of a speech frame. On the other hand, the replay speech signal, $r_e[n]$, can be modeled as the convolution of the natural speech signal, $s[n]$, and the impulse response of the intermediate devices, $h_1[n]$, (playback and recording device) along with propagating acoustic environment, and it is given by [175]:

$$r_e[n] = s[n] * h_1[n], \tag{3.18}$$

where $h_1[n]$ have the extra convolved components due to replay. The $h_1[n]$ is lump together with the impulse responses of recording device, playback device, and environment as given by Eq. 3.19. The genuine speech signal, $s[n]$, is inevitably convolved with the impulse response of the transmitting loudspeaker, $h_{mic}[n]$, before further convolution with the response of the channel under test, $h_1[n]$. The replay signal $r_e[n]$ is the channel output after further modification by the receiver response. This is, in fact, an approximation. In order to estimate the channel impulse response alone, the combined effects of the genuine speech, the transmitting loudspeaker, and the receiver must be removed from the replay signal using deconvolution. In particular, it is the combination of impulse responses of recording device, $h_{mic}[n]$, recording environment, $a[n]$, playback device (speaker), $h_{spk}[n]$, and playback environment, $b[n]$, i.e., $h_1[n]$ *lump* together all the three impulse responses.

$$h_1[n] = h_{mic}[n] * a[n] * h_{spk}[n] * b[n]. \tag{3.19}$$

In addition, if the presence of extra additive noise, $\eta[n]$, is added along with the convolution of impulse response it is more complex, then Eq. (3.18) becomes,

$$r_e[n] = s[n] * h_1[n] + \eta[n]. \tag{3.20}$$

To sum up, we assume the impulse response of microphone and loudspeaker as output of *linear* systems.

The speech signal recorded with the playback device contains the convolutional and additive distortions from the intermediate devices. The most crucial part in the detection of replay attack is during the process of feature extraction. To obtain the discriminatory information of genuine and replay speech signal, the focus should be on the representation of the spectral characteristics obtained from the intermediate devices. Eq. (3.18) represents the convolution term that transforms to the additive relation when converted to the real cepstral-domain (by ignoring phase information), and it is given by [176]:

$$r_e = \mathbf{s} + h_1, \tag{3.21}$$

where r_e, $\mathbf{s}$, and h_1 represents the cepstral vectors of replay, genuine speech signal, and the impulse response of intermediate devices, respectively. The features obtained from the vector h_1 can be used by subtracting the cepstral vector of genuine speech signal from that of replay speech signal. The features extracted from replay signal are also affected by the recording process.

The acoustical behavior of the speech signal recorded in different environment have differences in the speech signal. The speech signal when recorded in noisy environment will have distortion in the signal. However, its effect on the acoustical characterization of replay is yet to be analyzed. Replay speech is affected by the reverberation, which is included during recording of the speech signal and hence, basics of reverberation effect is explained in the next Section 3.6.2.

3.6.2 Basics of Reverberation

The replay speech signal is the re-recording of the target speaker's voice captured unknowingly from a distance with the help of a recording device. The recording can be done at different places, such as bedroom, balcony, canteen, office, home. When the recording is done particularly within the closed room, the reverberation is introduced severely during replay mechanism. Reverberation is the phenomenon to resist the sound after it has been stopped as a result of multiple reflections from surfaces, such as furniture, people, air medium. within a closed surface [177]. These reflections build up with each reflection and decay gradually as they are absorbed by the surfaces of objects in the space enclosed. The reflections are 1^{st} order (with only one deviation), and 2^{nd} order (with two deviations) from the wall, surface, and direct path without any deviations. The impulse response is known to carry the information of the acoustic environment, however, under assumption of Linear Time-Invariant (LTI) system [25, 178, 179]. Conventionally, replay signal (with reverberation), $s_{rev}[n]$, is modeled as a convolution of the natural speech signal, $s[n]$, with impulse response of acoustic environment, $r[n]$, [180, 181], i.e.,

$$s_{rev}[n] = s[n] * r[n]. \tag{3.22}$$

If the additive noise $\eta[n]$ is present then Eq. (3.22) becomes

$$s_{rev}[n] = s[n] * r[n] + \eta[n]. \tag{3.23}$$

The natural speech is repeated, time-shifted, and scaled for every non-zero point in the impulse response and the resulting signals are summed as shown via a schematic representation in Figure 3.11. The non-stationary monocomponent sig-

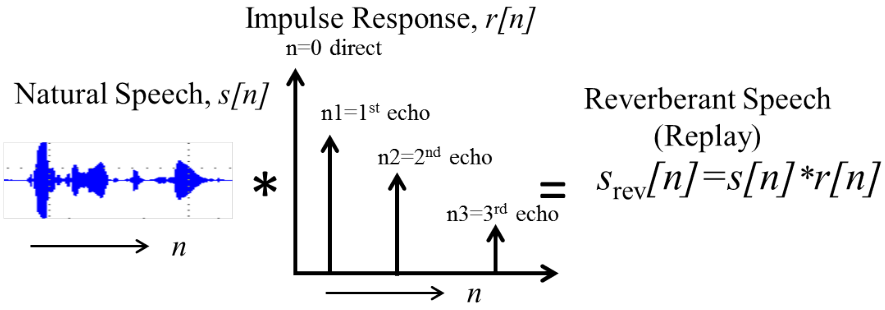


Figure 3.11: Convolution of Genuine Speech with Impulse Response (i.e., Sequence of Impulses at Different Echo Locations) in Order to Obtain the Reverberant Replay Speech Signal. After [9].

nals can be mathematically expressed as [182,183]:

$$s[n] = a[n] \cos \phi[n], \tag{3.24}$$

where $a[n]$ is the slowly-varying instantaneous amplitude, and $\phi[n]$ is the instantaneous phase [183]. The non-stationary multicomponent signal can be defined as the superposition of M monocomponent signals given as:

$$s_{multicompnent}[n] = \sum_{i=1}^{M} a_i[n] \cos \phi_i[n]. \tag{3.25}$$

As discussed earlier, reverberation includes the delay and change in amplitude forming the close copies of genuine signal that corresponds to different reflections [184]. Modeling the reverberation and understanding how the parameters related to the model affect both physical and perceptual properties of reverberation. Different types of reverberation models are time-frequency room model, novel signal-based measurement, reverberation decay tail measure, colouration measure. From a signal processing viewpoint and under the assumption of a fixed acoustic environment, reverberation can be modelled as a linear time-invariant (LTI) system with room impulse response (RIR), $h[n]$, with the input signal, $s[n]$, to give the output signal $s_{rev}[n]$. The reverberation process can then be written as the convolution between the input and the RIR:

$$s_{rev}[n] = \sum_{i} k_i s[n - n_i], \tag{3.26}$$

where $s[n]$ and $s_{rev}[n]$ are the genuine and reverberated signals, respectively, and k_i and n_i are the change in amplitude and delay of each samples, respectively, for i reflections that occurred in the closed room. When we compare Eq. (3.25) and Eq. (3.26), we can say that reverberation changes monocomponent signal into multicomponent signals. The duplicates are spectrally very close to each other [184].

Reverberation introduces delay and attenuation to produce close copies of the genuine signal corresponding to the different reflections of the acoustical signal in the environment [184]. It can be observed from Figure 3.12 that the replay speech samples are shifted from the genuine components, and the amplitude also varies compared to the genuine signal.

Discrete early reflections (in particular, 1^{st} or 2^{nd} order reflections) are typically involved in the early regions of an impulse response. The discrete early reflec-

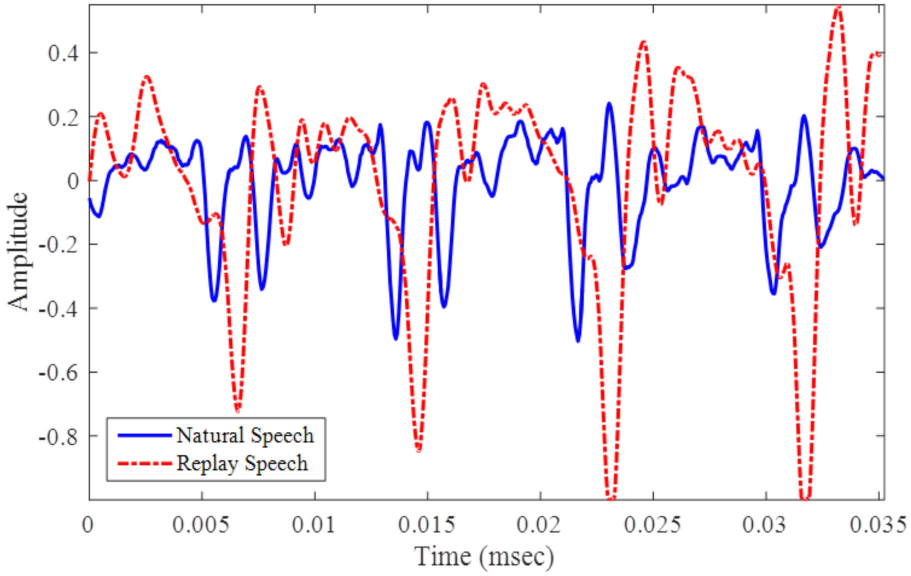


Figure 3.12: Segment of Speech Signal Showing the Effect of Reverberation for Replay Signal (Dotted Line) in Terms of Delay in Each Speech Sample, and Changes in Amplitude Compared to the Genuine Speech Segment (Solid Line). After [9].

tions can be simulated by means of a tapped delay line, which allows replicating some versions of the input signal, each delayed in a different amount [185]. The time-domain speech signal are shown for both genuine (Figure 3.13(a)), and reverberated speech signal in (Figure 3.13(b)). The reflections further become densely packed in time-domain, composing the diffuse tail (as seen in Figure 3.13(b)) [185]. The time of the peak indicates, 'how long the reflected signal will arrive at the recording device?', and the amplitude of the peak shows the amplitude of the reflected signal [185]. The first peak of the reverberated signal corresponds to the signal that arrives directly from the source of the recording, which arrives with the shortest possible delay. The other subsequent peaks arrives because of reflections, each related to its particular path that come in its way. Eventually, the reflections become sufficiently dense that they indeed overlap in time. Because energy is absorbed by environmental surfaces with each reflection (as well as by air), longer paths produce lower amplitudes, and the overlapping echoes produce a “tail" in the impulse response that decays with time [185].

If a room does not have any signal absorbing surfaces, such as wall, roof, and floor, the signal bounce back between the surfaces, and takes very long (ideally infinite) time for the signal to end. In such a room, the listener or the recording device will hear/record both the direct signal as well as the repeated reflected signal waves. If these reverberations are more excessive, the sound will run together with a mere loss of articulation, and it becomes *muddy* and also *garbled* [177]. The larger rooms have few reflections resulting in slow decay of reverberated signals,

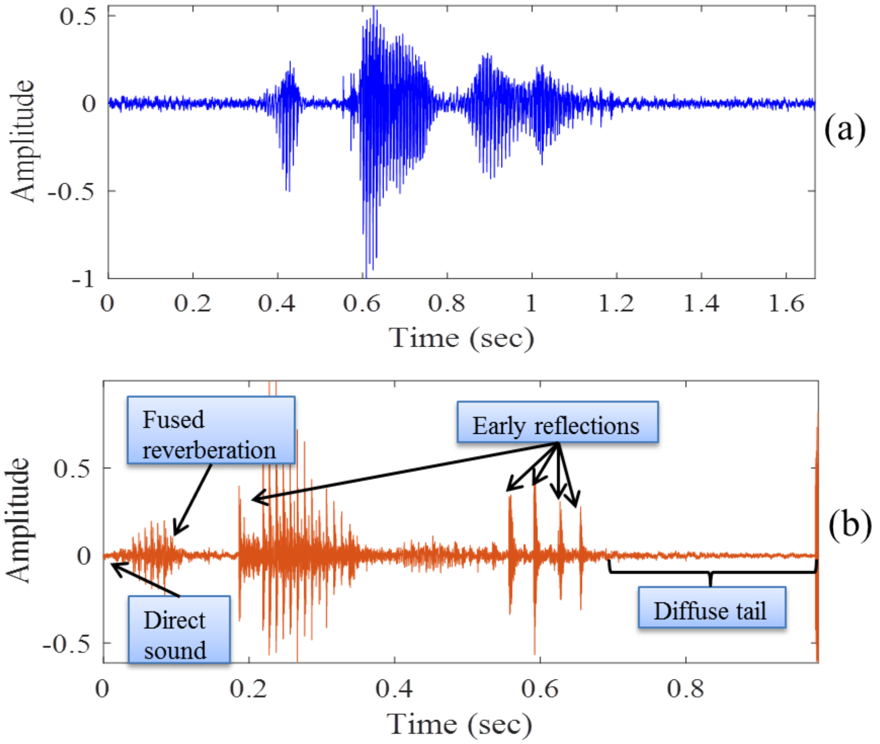


Figure 3.13: Time-Domain Speech Signal for (a) Genuine, and (b) Reverberated (Replay). After [9,25].

and the decay rates are also affected by material, such as carpet, curtains, sofa-sets. Reverberation is also found to distort the structure of source signals in the spectral energy density as shown in Figure 3.14 via spectrogram [14,27,60,185]. The time-domain speech signals for genuine (Panel I), and replay speech (Panel II) are shown in Figure 3.14(a) corresponding to their spectral energies in Figure 3.14(b). The highlighted regions in spectral energy densities show the distortion that are included because of reverberation. Distortion is due to *decay* in spectrum and hence, a kind of *energy loss* [186].

TEO is applied on filtered subband signals that are obtained from the Gabor filterbank, and estimate instantaneous energy profiles for each narrowband signal. The Teager energy profiles are passed through frame-blocking with window length of 20 ms and shift of 10 ms followed by logarithm operation. This subband Teager energy spectrum shows the difference between natural and replay speech signal.

The spectral energy density obtained via traditional spectrogram, and Teager energy spectral features shown in Figure 3.15(b) and Figure 3.15(c). The Panel I and Panel II corresponds to the spectral energy density for natural and replay speech signal, respectively. From Figure 3.15, it can be observed that the spectral energy densities obtained from the Teager energy gave high energies in both

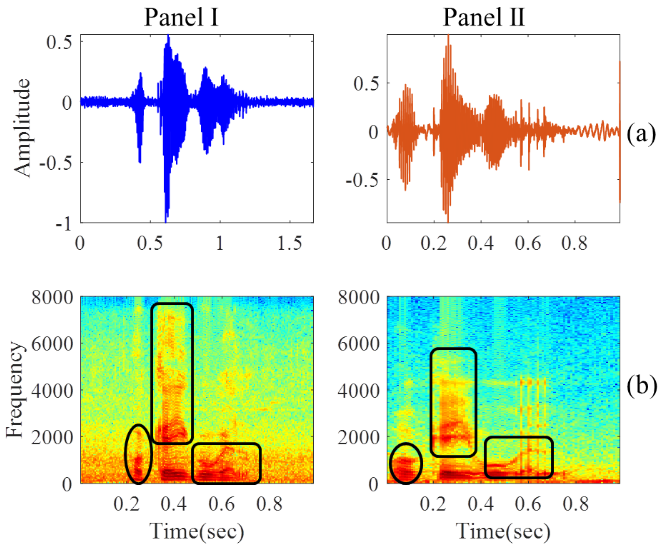


Figure 3.14: (a) Time-Domain Speech Signal, and (b) Corresponding Spectral Energy Densities via Spectrogram of Genuine (Panel I), and Replay (Panel II) Speech Signal, Highlighted Regions Oval, and Boxes Show the Distorted Spectral Regions. After [9].

lower and higher frequencies as compared to the ones shown by the traditional spectrogram. The similar pattern was observed for the replay speech signal (Panel II). Highlighted regions in Figure 3.15 shows the energy differences corresponding to the natural and replay signals. These energies obtained from the proposed approach contributes to detect replay signal.

Further simulation is performed to observe the effect of reverberation on the Teager energy profiles of synthetic speech (i.e., simulated genuine), and corresponding replay signals in Figure 3.16. The train of impulses (Figure 3.16(a)) is convolved with a damped sinusoid signal (Figure 3.16(b)) producing a convolved signal (Figure 3.16(c)). Now, assume the convolved signal in Figure 3.16(c)) is a simulated genuine speech signal. Now, to obtain a reverberated signal, we convolved the simulated genuine speech signal (Figure 3.16(e)) with a train of impulses representing echo (Figure 3.16(d)), and obtained a synthetic reverberated signal having close copies of original genuine signal (Figure 3.16(f)). The impulse response for reverberated signal in Figure 3.16 had an echo kept with a particular time interval, i.e., impulse arrives at every 8 ms. However, in real case scenario, it may not be the case, i.e., the echo impulses may arrive with a small interval gap or it may arrive with a large delay as well depending upon shape and size of the acoustic environment.

We observed that the Teager energy traces of replay speech signal segment recorded for different environments, such as (Panel I) balcony, (Panel II) bedroom, (Panel III) canteen, and (Panel IV) office are as shown in Figure 3.17(b). The extra

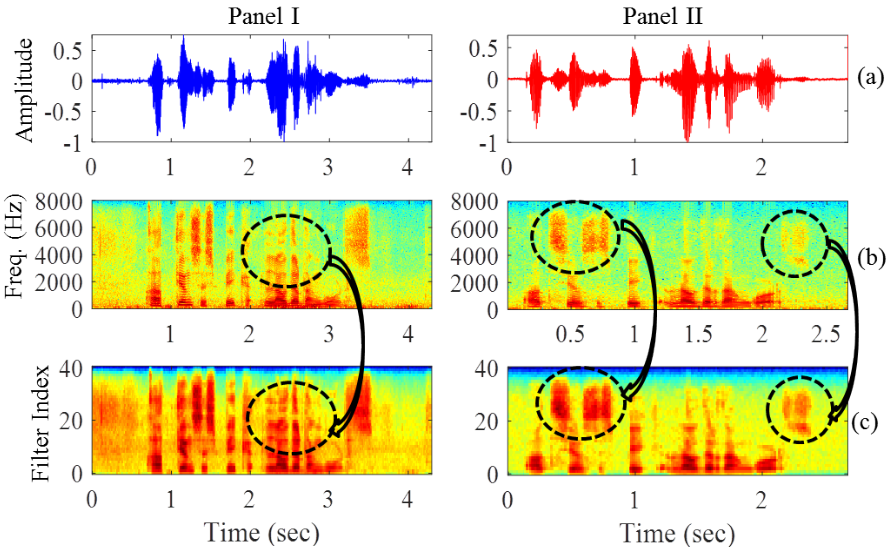


Figure 3.15: (a) Time-Domain Speech Signals, Comparison of Spectral Energy Density via (a) Spectrogram, and (b) Teager Energy-Based Approach. Panel I and Panel II Shows Natural and Corresponding Replay Speech, respectively. Highlighted Oval Regions via Ellipses Shows the Difference in Pattern of Spectral Energy Densities. After [9].

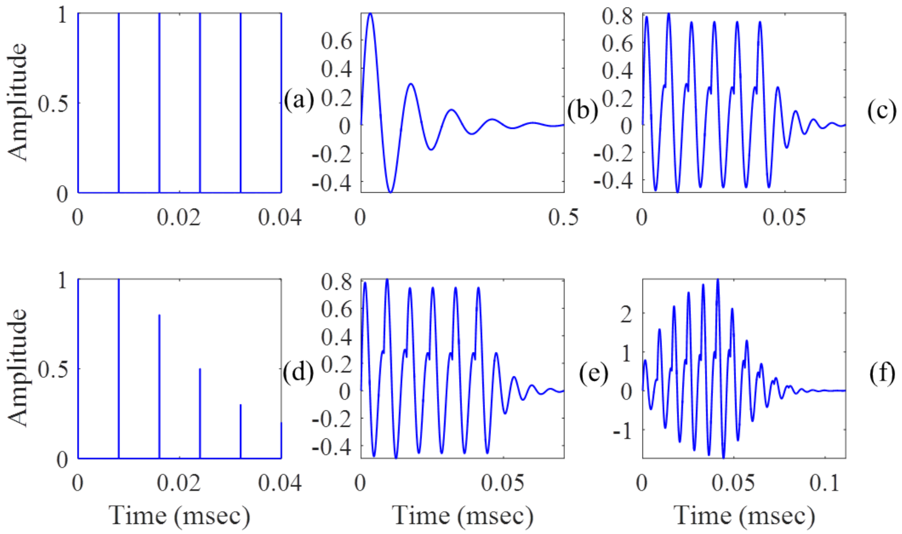


Figure 3.16: (a-d) Train of Impulses and Echoes to Model Reverberation; (b) Damped Sinusoid Signal; (c-e) Convolved Signal Obtained from (a) and (b); and (f) Convolved Signal Obtained from (d) and (e). After [9].

pulses are observed when the replay speech signal is recorded in a closed room, such as bedroom and office as shown in Figure 3.17 (Panel II and Panel IV (b)). On the other hand, extra impulse-like energy traces are not observed for replay speech recorded in balcony, and canteen environment Figure 3.17 (Panel I and Panel III (b)).

The TEO profiles show high energy pulses around the Glottal Closure Instant (GCI), because of impulse-like excitation to vocal tract system and this sudden

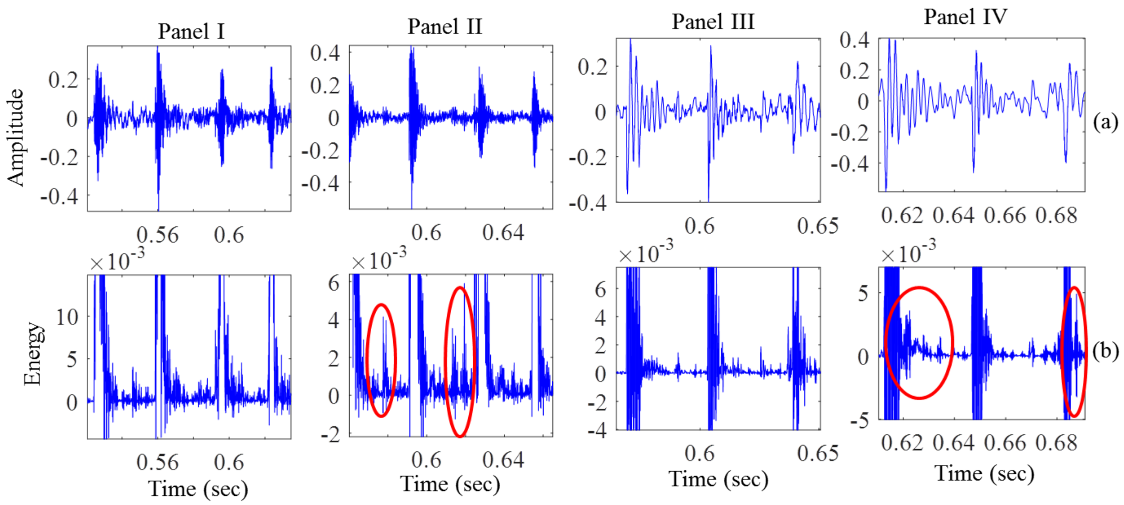


Figure 3.17: (a) Time-Domain Speech Segment of Replay Signal Recorded in Balcony (Panel I), Bedroom (Panel II), Canteen (Panel III), and Office (Panel IV) Along with Their Corresponding Teager Energy Profiles (b). Highlighted Ovals Show the Extra Impulse-Like Teager Energy Traces Observed Replay Speech Recorded in Closed Room.

glottal closure produces high energy (also observed in Section 3.3) and thus, TEO produces high energy around these regions [168]. Along with high Teager energy pulses, the bumps are observed around the energy pulses, indicating significant contribution of nonlinear effects during the speech production process [168] (please refer Section 3.3 for more details).

3.7 Experimental Setup

In this Section, we provide the details of SSD database used to perform the experiments on the proposed feature sets. The performance of SSD system is evaluated on various standard databases (as discussed in Chapter 2). In addition, the feature extraction parameters, and the model training information is also discussed in this sub-Section.

3.7.1 Feature Parameterization

The details of the parameters used for feature extraction of various feature sets is explained in Table 3.1.

3.7.2 Model Training and Score-Level Fusion

We used the Gaussian Mixture Model GMM with 128 mixtures for modeling the two classes, in which the classes correspond to the genuine and bonafide class in ASVspoof 2015 database. The GMMs are trained with the training set of the

Table 3.1: Details of Features Extraction Parameters. After [9]

Parameters	CQCC	LFCC	MFCC	TECC
Frequency Scale	-	Linear	Mel	Linear
Subband Filters	-	40	40	40
Fmin (in Hz)	15	0	0	10
Fmax (in Hz)	8000	8000	8000	8000
Dimension of a Feature Vector	90	120	39	120

database. The use of a GMM classifier comparatively perform well for the detection of genuine *vs.* bonafide speech in the ASVspoof 2015 challenge [187]. Final scores are represented in terms of the log-likelihood ratio (LLR). The decision of the test speech being genuine or bonafide is based on the LLR, i.e.,

$$LLR = \log \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_1)}, \tag{3.27}$$

where $p(\mathbf{X}|H_0)$, and $p(\mathbf{X}|H_1)$ are the likelihood scores from the GMM for the genuine and bonafide trials (with hypothesis H_0 and H_1), respectively, for feature vectors, $\mathbf{X}$. To obtain the complementary information of the MFCC, and TECC feature sets, we use their score-level fusion, i.e.,

$$LLR_{combine} = (1-\alpha)LLR_{feature1} + \alpha LLR_{feature2}, \tag{3.28}$$

where $LLR_{feature1}$ is the log-likelihood score of MFCC, and $LLR_{feature2}$ is the score for TECC, respectively. The weights of the scores are decided by the fusion parameter, α. We compared TECC results with other state-of-the-art features sets, such as Mel Frequency Cepstral Coefficients (MFCC) [13], Constant Q Cepstral Coefficients (CQCC) [110, 112], and Cochlear Filter Cepstral Coefficients-Instantaneous Frequency (CFCC-IF) [111].

3.8 Experimental Results of the SSD Task

3.8.1 Results on ASVspoof ASVspoof 2015 Challenge

3.8.1.1 Results on Development Set

The results obtained in % Equal Error Rate (EER) of TECC feature set on development and evaluation sets are shown in Table 3.2. From the results, it can be observed that on development set, the proposed feature set has much less % EER of 0.38 % compared to the CFCC-IF, and MFCC. However, the best performing feature set, i.e., CQCC gave lower % EER of 0.038 %. We further used score-level fusion of MFCC and TECC feature sets in order to obtain possible complementary information, and further reduce the % EER on both development and evaluation sets. However, we could not obtain the reduced % EER.

Table 3.2: Comparison of Results in % EER ASVspoof 2015 Challenge Database. After [10]

Feature Set	Development	Evaluation
MFCC [13]	6.14	9.15
TECC	0.38	5.95
CFCC-IF [111]	2.29	1.211
CQCC [110]	**0.0381**	**0.255**
TECC+MFCC	0.38	6.41

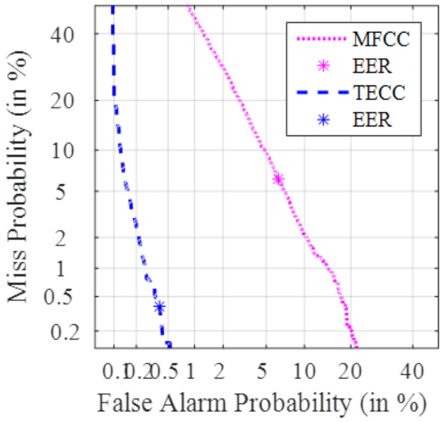


Figure 3.18: Individual DET Curves of TECC, and MFCC Feature Set on Development Dataset of ASVspoof 2015 Challenge Database. After [10].

The performance is also shown in Fig. 3.18 by the Detection Error Trade-off (DET) curve on development set for MFCC and TECC feature sets. It can be observed from the DET curve that the operating points obtained from the score of MFCC have high miss probabilities and false alarm, whereas TECC feature set has a significantly lower false alarm, and miss probabilities in the DET curve.

3.8.1.2 Results on the Evaluation Set

On evaluation set, the dataset is divided into two groups, namely, known (S1-S5), and unknown attacks (S6-S10). The unknown attacks were included during the challenge, which are not used in the training and development datasets. These unknown attacks are challenging to detect, in particular, the S10 attack which is

Table 3.3: Results in % EER on Evaluation Dataset for Each Spoofing Attack. Both Known and Unknown Attacks. +:Score-Level Fusion. After [10].

Feature Set	Known Attacks						Unknown Attacks						All Avg.	S1-S9 Avg.
	S1	S2	S3	S4	S5	Avg.	S6	S7	S8	S9	S10	Avg.		
MFCC	2.34	9.57	0.00	0.00	9.01	4.18	7.73	4.42	0.3	5.17	52.99	14.12	9.15	4.28
TECC	0.00	0.21	0.00	0.16	0.67	0.20	0.41	0.00	0.00	0.00	58.14	11.71	5.95	**0.161**
CFCC-IF	0.101	0.863	0.000	0.000	1.075	0.408	0.846	0.242	0.142	0.346	8.490	2.013	1.211	0.39
CQCC	0.005	0.106	0.000	0.000	0.130	0.048	0.098	0.064	1.033	0.053	1.065	0.462	0.255	0.163

developed using Unit Selection Synthesis (USS)-based approach. The detailed % EER of MFCC, TECC, CFCC-IF, and CQCC on both known and unknown attacks are reported in Table 3.3. It can be observed that for spoofing attacks (S1 to S9), for most of the cases, TECC feature set gave lower % EER compared to the other state-of-the-art feature sets. For known attacks, the average % EER of TECC is 0.20 %, and the average % EER for unknown attacks (S6-S9) is 0.161 %, which is lower compared to the other feature sets. The comparison in % EER from S1 to S9 spoofing algorithms are shown in Fig. 3.19. We can observe that the CFCC-IF feature set has higher % EER (green dotted line) compared to the CQCC and TECC feature sets. For individual spoofing attacks of S7, S8, and S9, it can be observed that the % EER is equal to 0 %, which is best performing system than the CQCC feature set. However, the TECC feature set fails to detect the S10 (USS) spoof speech signals resulting in higher % EER of 58.14 % that increases % EER for entire SSD task. This may be due to the fact that USS-based spoof contains concatenation of natural speech sounds units resulting in similar bumps in TEO profiles w.r.t nonlinearity in speech production and thus, creating a larger confusion during SS *vs.* natural SSD task.

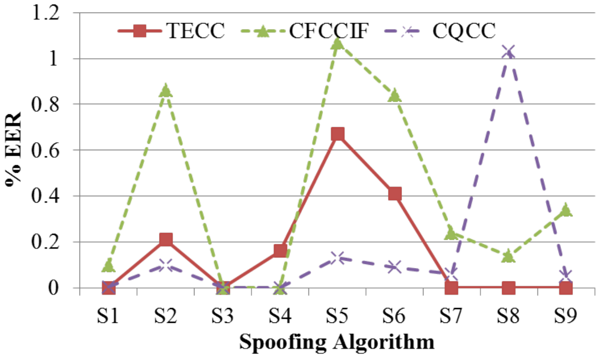


Figure 3.19: Comparison of S1-S9 Spoofing Algorithms in % EER of CFCC-IF (Green Line), CQCC (Purple Line), and TECC Feature Set (Red Line). After [10].

3.8.2 BTAS 2016 Database

The results obtained in EER of TECC feature set on development and evaluation sets are shown in Table 3.4. We compared our results with the baseline system, MFCC, and CQCC feature sets. From the experimental results, it can be observed that the TECC feature set has much more less EER of 2.25 %, and 4.51 % on dev and eval set, respectively, compared to the baseline system, MFCC, and CQCC feature sets.

Table 3.4: Results (in % EER) for BTAS 2016 Database. After [10]

Subset	Baseline	MFCC	CQCC	TECC
Dev	5.91	3.66	3.05	2.25
Eval	-	7.59	18.86	4.51
Fusion with TECC				
Dev	-	**2.20**	2.25	-
Eval	-	**4.43**	4.50	-

We further used score-level fusion of MFCC and CQCC with TECC feature set in order to obtain possible complementary information, and reduce the % EER further on both development and evaluation sets (as shown in Table 3.4). The score-level fusion reduced the % EER to 2.20 % with MFCC and TECC feature sets (with fusion factor, α= 0.8) and with CQCC feature set, it reduced to 2.31 % (with fusion factor, α= 0.9). On the other hand, on evaluation set, the score-level fusion reduced only fusion of MFCC and TECC and gave % EER of 4.43 % (with fusion factor, α= 0.9), whereas with CQCC feature set, the % EER did not reduce.

Table 3.5 shows the performance on evaluation set in % HTER on baseline system, MFCC, CQCC, and TECC feature sets. It can be observed that TECC feature set gave lower % HTER compared to the other feature sets. Furthermore, we analyzed individual presentation attack (as reported in Table 3.5). In the Table 3.5, 'SS' stands for speech synthesis, 'VC' stands for voice conversion, 'RE' stands for replay, 'LP' stands for laptop, 'PH1' is Samsung Galaxy S4 phone, 'PH2' is iphone 3gs, 'PH3' is iphone 6s, and 'HQ' stands for high quality speakers were used during replay attack. It can be observed that for all the types of replay attacks, we obtained lower % HTER with proposed TECC feature set. However, for unknown attacks (highlighted with bold font), we obtained higher % HTER for all the feature sets, which means degradation in the overall performance.

The histogram plots of log-likelihood scores obtained from the Gaussian mixtures corresponding to (a) MFCC, (b) CQCC, and (c) TECC are shown in Figure 3.20 for development (Panel I), and evaluation set (Panel II), respectively. It can be

Table 3.5: Individual Attack Results (in % HTER) for Eval Set. After [10]

Attacks	Baseline	MFCC	CQCC	TECC
SS-LP-LP	2.87	10.82	50	**2.39**
SS-LP-HQ-LP	2.87	14.89	50	**1.75**
VC-LP-LP	3.58	4.05	50	**1.43**
VC-LP-HQ-LP	3.39	3.99	50	**1.32**
RE-LP-LP	17.02	9.40	50	**1.77**
RE-LP-HQ-LP	11.24	28.25	50	**3.02**
RE-PH1-LP	52.24	29.37	50	**24.77**
RE-PH2-LP	51.96	**27.65**	50	29.87
RE-PH2-PH3	51.56	**38.85**	50	50.17
RE-LPPH2-PH3	**20.62**	47.87	50	41.92
Average	6.87	6.89	50	**3.71**

Bold font indicates they are unknown attacks

observed that for TECC feature sets, the LLR scores of both natural and spoof are properly distributed resulting in less % EER as compared to the distribution corresponding to other feature sets on development set. Similar observation is found on evaluation set for MFCC, and TECC feature sets. From the Figure 3.20, it can be observed that huge change in score distributions on development (i.e., -10 to 10), and evaluation (i.e., -80 to -10) sets for CQCC feature set. This in turn results in high % HTER for CQCC on evaluation set, as HTER depends on the threshold of development set (which is near to 0 (Figure 3.20 Panel I (b))).

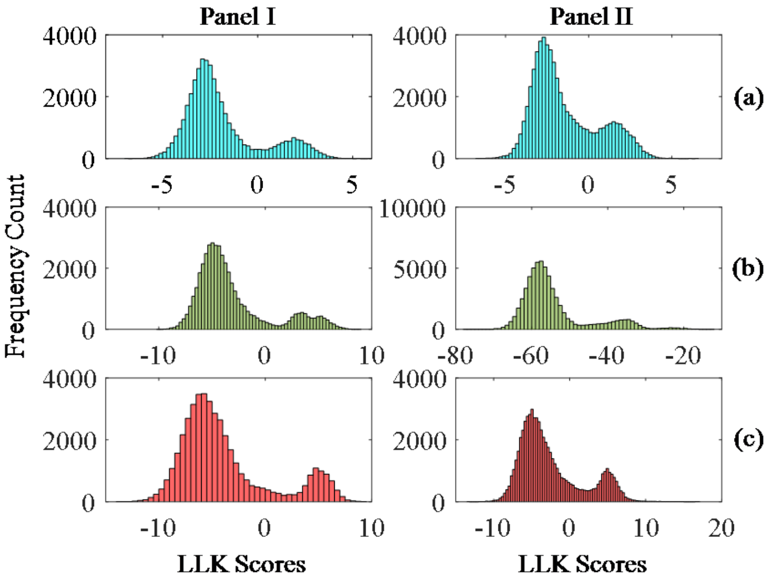


Figure 3.20: Histogram Plots for Panel I: Development, and Panel II Evaluation Set. (a) Score Distribution of MFCC, (b) CQCC, and (c) TECC Feature Set. After [10].

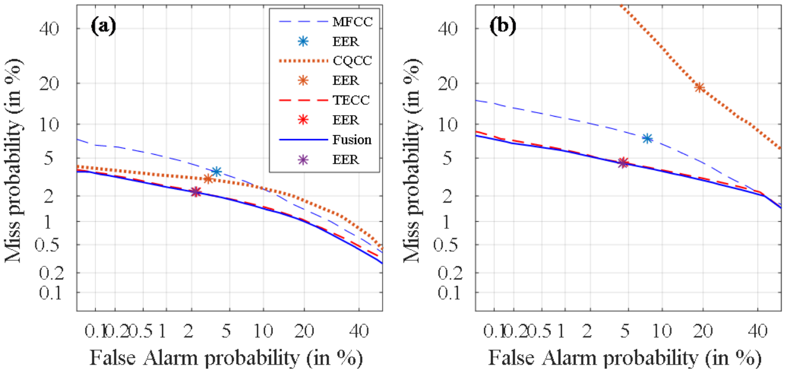


Figure 3.21: DET Curve for (a) Development, and (b) Evaluation Set. After [10].

The performance is also shown with DET curves for all the feature sets along with their best score-level fusion in Figure 3.21(a), and Figure 3.21(b). From Figure 3.21(a), it can be observed that for MFCC, and CQCC shows high miss probability, and false alarm probability, respectively, which is not a good case for the voice biometric or ASV system. However, the TECC feature set along with score-level fusion of CQCC and TECC feature set shows the reduced miss probability and false alarm probability compared to the other feature sets. On the other hand, for evaluation set, the DET curves for all the feature sets have high probability with high false alarm rate, which indicate that the evaluation set is very challenging to develop a suitable countermeasure.

3.8.3 Results on ASVspoof 2017 Challenge v2.0

3.8.3.1 Results on Development Set

This Section presents the experiments performed on development set in order to optimize the parameters on the evaluation set, such as the approximate subband filtered signals, and the choice of bandwidth of subband filter in Gabor filterbank.

Effect of Subband Filters

The human auditory system carries several thousands of subband filters which results in a dense filterbank in frequency-domain [53], [166, 167]. To estimate the Teager energy features accurately, we increased the number of subband filters in Gabor filterbank. Results with increase in the number of subband filtered signals are shown in Figure 3.22. It can be observed that with 40 number of subband filters in filterbank, we obtain very high EER of 25.07 % on development set. As we increase the number of subband filters in filterbank, the EER goes on decreasing from the EER obtained using 40 number of subband filters. The low EER of 11.82

% was obtained with 80 number of subband filtered signals. However, when we further increase the subband filtered signals to 100, the EER increases. The subband filters overlap with each other and hence, discriminative information is lost that results to degrade the SSD performance.

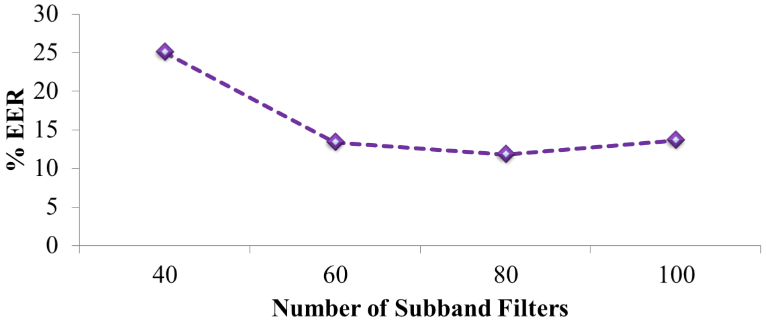


Figure 3.22: Results in % EER on Development Set of ASVspoof 2017 Challenge Database With Varying The Number of Subband Filtered Signals in a Filterbank. After [9].

Effect of Bandwidth of Subband Filters

The formant transitions, in particular, the higher formants are important when it comes to speaker-related information (namely, speaker identification or verification task). The higher formants or the energy present in higher frequency indeed help to detect the replay speech signal from its natural counterpart. The higher spectral energy information depends on the process of how it is extracted, in particular, frequency scale used in filterbank, bandwidth of a subband filter, shape of subband filter. The choice of the bandwidth in a subband filter should not be much narrow neither it should be wider. If the bandwidth of the subband filter is too small then the filtered signal may not capture the formant transition well, whereas if the bandwidth is too large, the features extracted might be inaccurate [167]. Hence, after a certain bandwidth, further widening of the bandwidth result in poor frequency resolution and hence, it degrades the performance [27], [166, 167]. Using 80 number of subband filters as it gave lower EER (discussed in Section 3.8.3.1), we performed further experiments by varying the bandwidth from 50 Hz to 400 Hz of a subband filter. The results obtained by varying the bandwidth are shown in Figure 3.23. Using 100 Hz bandwidth, we obtained lower EER of 10.80 % on development set compared to the other choices of bandwidths.

In addition, we also performed an experiment without integrating Gabor filterbank to investigate the importance of filterbank for the proposed approach.

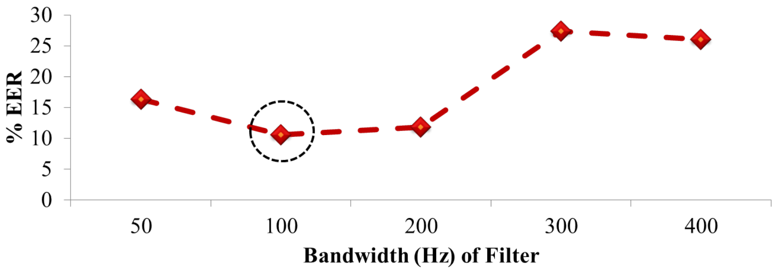


Figure 3.23: Results With Varying the Bandwidth of a Subband Filter on Development Set of ASVspoof 2017 Challenge Database. After [9].

Table 3.6 shows the results (with and without filterbank), on development and evaluation sets for TECC feature sets. It can be observed that the EER obtained for TECC feature set gave high EER for both development and evaluation sets.

Table 3.6: Results on ASVspoof 2017 Database Obtained With and Without Applying Filterbank. After [9]

Feature Set	Without Filterbank		With Filterbank	
	Dev	Eval	Dev	Eval
TECC	40.94	42.33	**10.80**	**11.41**

3.8.3.2 Results on Evaluation Set

Based on the experiment performed on development set, parameters are optimized on development set, and later carried forward on the evaluation set. In particular, 80 number of subband filters using 100 Hz bandwidth of a subband filter in filterbank is used to extract the TECC feature set. In addition, we analyzed the effect of EER depending on replay configurations, in particular, different acoustic environment, playback, and recording devices on the evaluation set.

In addition, we also extended our experimentation to test the accuracy on deep learning models. From the results, it was observed that there was a big reduction in EER for evaluation dataset for GMM-based model as compared to the Convolutional Neural Network (CNN) model indicating that the proposed TECC feature set being cepstral feature is not performing well when used with DNN as classifier. In similar line, in speech recognition literature, Mel filterbank energy features were found to be effective for DNN-based than MFCC-GMM-HMM systems. The CNN architecture consists of convolutional, pooling, and fully-connected (FC) layers. It is a type of Deep Neural Network which uses convolutional layers and pooling layers in order to extract abstract feature data, followed by FC layers. Batch Normalization is also used in order to prevent overfitting. The result for the experiments performed with neural network-based classifier is reported in Table 3.7. The TECC feature set do not perform well for the neural network-based classifier compared to the GMM-based classifier resulting in lower EER of 10.0 % and 11.41 % on dev and eval set, respectively.

Table 3.7: Results (in % EER) with GMM and CNN Classifier on TECC feature sets. After [11].

Classifier	Dev	Eval
GMM	10.80	11.41
CNN	24.84	29.49

Results using Score-Level Fusion

Table 3.8 show the results in EER on development and evaluation sets. We compared TECC feature set with the other existing feature sets, namely, CQCC, MFCC, and LFCC. On ASVspoof 2017 version 2.0 database, the post evaluation baseline is modified from the earlier baseline in the form of having the log-energy coefficients, and Cepstral Mean Variance Normalization (CMVN) method, the enhanced baseline results are reported in Table 3.8. However, we considered the CQCC feature set with CMVN method as our first baseline to have a fair comparison, as the other feature sets are in the cepstral-domain. In addition, we also considered LFCC as our second baseline, since TECC feature set is extracted with the linear frequency scale. The MFCC feature set is also compared as it is one of the state-of-the-art feature set used in the speech literature. From Table 3.8, it can be observed that the relatively low EER obtained is with TECC feature set resulting in 10.80 % and 11.41 % on development and evaluation sets, respectively.

Furthermore, in order to increase the performance of the replay SSD task, we further performed score-level fusion as per Eq. (6.7) to obtain possible complementary information. The low EER obtained is with score-level fusion of TECC and LFCC feature sets that resulted in 8.10 % and 10.49 % EER at fusion weight of $\alpha = 0.7$ on development and evaluation sets, respectively. (Please note that the performance on evaluation set was not done using oracle fusion (If 'a' is optimized on the evaluation data, such fusion should be called oracle fusion)). In addition, we also fused the scores of all the feature sets used and observed the importance of TECC feature set. It can be observed from the Table 3.8 with the score-level fusion of TECC along with CQCC, MFCC, and LFCC the performance of replay detection is better compared to the other fusion of feature set, indicating that the proposed feature set captures *complementary* information than the other

Table 3.8: The Final Results (in % EER) on Development and Evaluation Sets. After [9]

Feature Set	Dev	Eval
CQCC (Baseline system)	9.06	13.74
CQCC (Our baseline1)	12.81	19.04
MFCC	24.19	26.90
LFCC (Our baseline2)	16.76	13.90
TECC	**10.80**	**11.41**
TECC+CQCC	8.90	11.77
TECC+MFCC	13.13	13.64
TECC+LFCC	8.10	10.49
CQCC+LFCC+MFCC	7.37	12.06
CQCC+LFCC+MFCC+TECC	**6.68**	**10.45**

feature sets alone or their fusion. The low EER obtained is 6.68 %, and 10.45 % on development and evaluation sets, respectively.

The performance evaluation is also shown by the DET curves for CQCC, MFCC, LFCC, and TECC feature sets along with their best score-level fusion results in Figure 3.24. It can be observed that the FRR of CQCC, MFCC, and LFCC is very high for the given FAR, which is not a good case for ASV system. There is significant decrease in miss probability fusing TECC feature set on development set as shown in Figure 3.24 (a), which is further reduced when the scores are fused with CQCC, MFCC, and LFCC feature set. We observe similar pattern of development set, on evaluation set also as shown in Figure 3.24 (b).

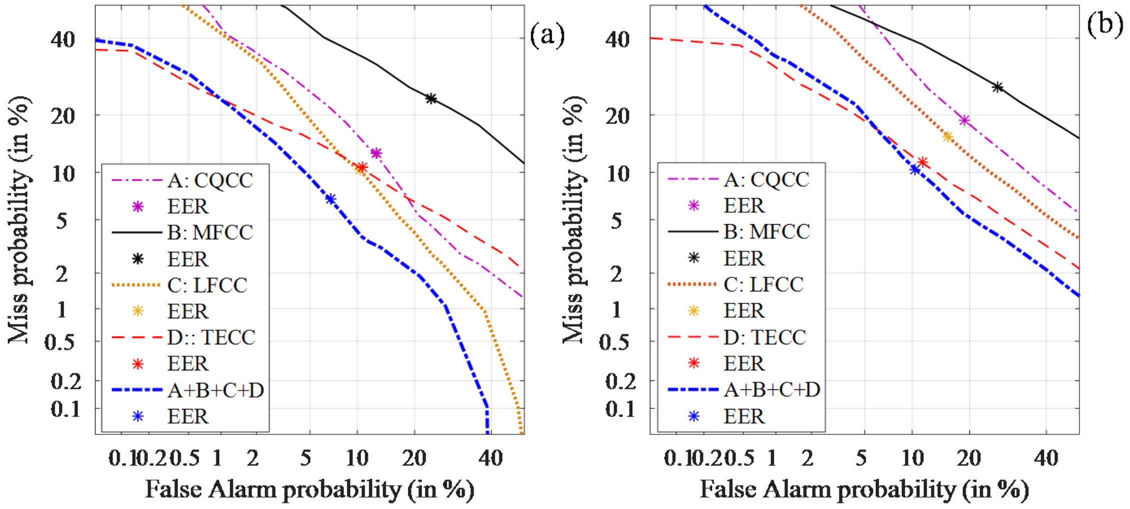


Figure 3.24: DET Curves for (a) Development, and (b) Evaluation Sets. After [9].

Effect of Replay Configurations (RC)

The updated ASVspoof 2017 challenge version 2.0 database provides the detailed description of replay configuration, in particular, acoustic environment, playback, and recording devices [6]. There are in total 61 distinct different replay configurations. The replay utterance encompass those of a playback and recording device along with an acoustic environment through which sound propagates [6]. On evaluation set, the EER with all the feature sets for different replay configurations are shown in Table 3.9. The overall performance of different replay configurations has least EER using TECC feature set. Hence, TECC feature set is able to detect different replay configurations better compared to the other feature sets. Furthermore, we analyzed the individual replay configurations discussed in the next sub-Section.

Table 3.9: Comparison of Feature Sets in % EER on Different Replay Configurations (RC). After [9]

Feature Set	Acoustic Environment	Playback Device	Recording Device
CQCC	17.85	16.43	18.06
MFCC	26.34	24.15	24.49
LFCC	15.18	14.48	14.85
TECC	**11.41**	**10.42**	**11.20**

The acoustic environment listed in [6] are the actual space in which the original speech data was re-recorded. The ASVspoof 2017 challenge version 2.0 database has in total 26 different environments denoted from E01-E26. Different environments have the variations included with the levels of additive ambient, convolutive, and reverberation noise. The level of noise in environment is assumed to be *inversely* proportional to the threat that pose to the ASV system. Figure 3.25 show the individual EER for various environmental conditions on evaluation dataset. We can observe that using MFCC and CQCC feature sets, the EER for most of the environments are relatively higher compared to the LFCC and TECC feature sets. However, TECC feature set show the lower EER for different environments.

Similar to different acoustical environments, there are 26 different playback devices denoted by P01-P26 [6]. The EER for all the different playback devices using all the feature sets are shown in Figure 3.26. Similar to the acoustical environments, TECC feature set gave relatively lower EER for different playback devices compared to the other feature sets.

There are 25 different recording devices used during collection of replay speech

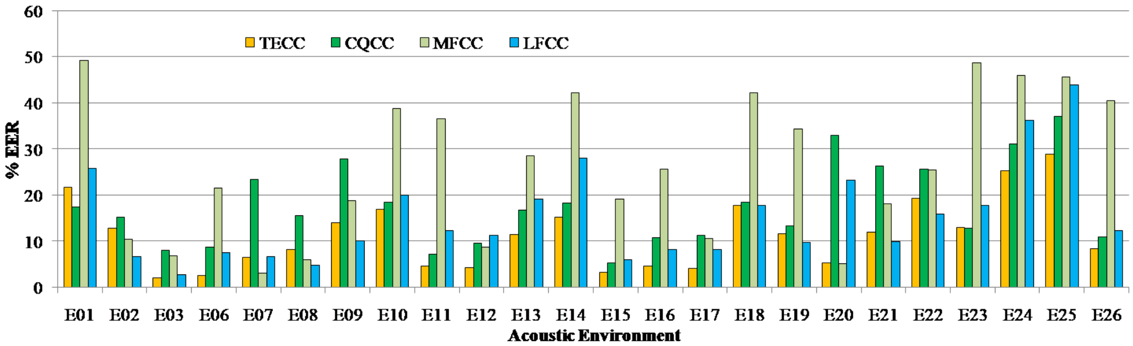


Figure 3.25: Individual % EER for Different Environment Conditions With CQCC, MFCC, LFCC, and TECC Feature Sets (Difference in % EER For All The Feature Sets is Highlighted by Oval). After [9].

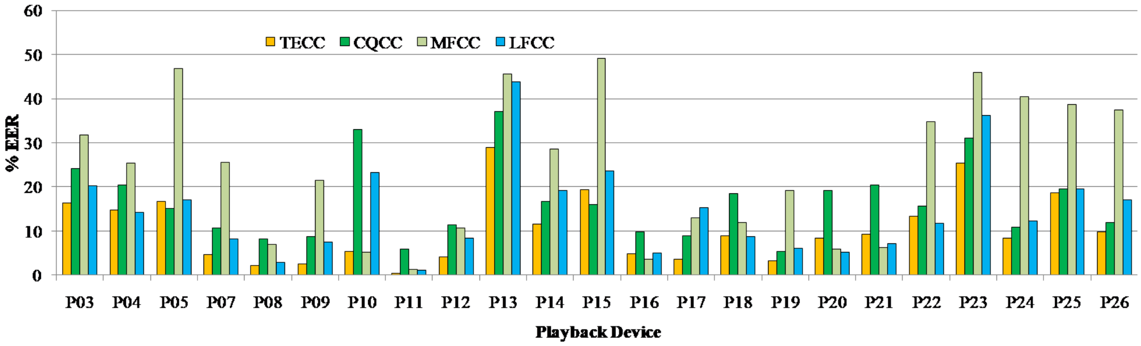


Figure 3.26: Individual % EER for Different Playback Devices With CQCC, MFCC, LFCC, and Proposed TECC Feature Sets (Difference in % EER for All the Feature Sets is Highlighted by Oval). After [9].

denoted by R01-R25 [6]. Figure 3.27 show the EER for different recording devices with all the feature sets on evaluation sets. Similar to acoustical environments and playback devices, the pattern of lower EER is observed with TECC feature set for different recording devices compared to the other feature sets.

The acoustic environment, playback devices, and recording devices are classified into three different levels of threat, namely, low, medium, and high. Figure 3.28 shows the EER for different levels of threat using CQCC, LFCC, MFCC, and TECC feature sets for all the different replay configurations. The high-level threats are difficult to detect because professional audio equipment, such as active studio monitors, and studio headphones were used to produce replay samples. In addition, samples collected from studio quality condenser microphones or handheld recorders are assumed to be of higher quality and hence, gives higher EER for high-level threats. As the level of threat goes on increasing, the EER also increases. The TECC feature set has lower EER for all the levels of threats compared to the other systems.

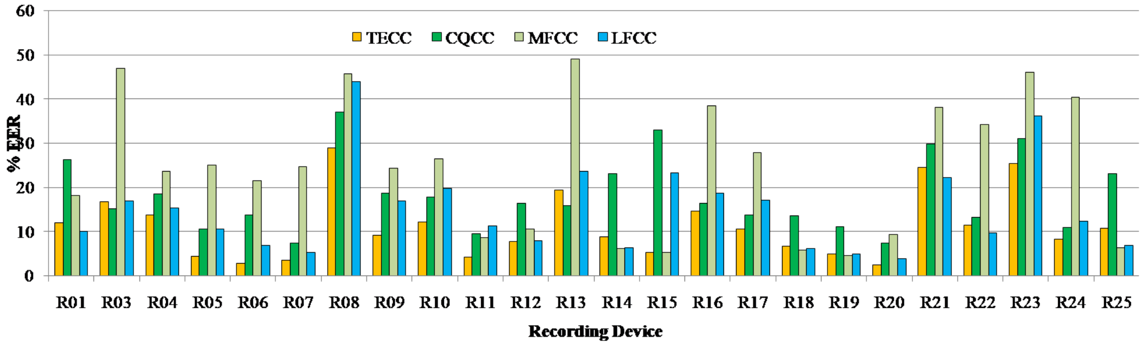


Figure 3.27: Individual % EER For Different Recording Devices with CQCC, MFCC, LFCC, and TECC Feature Sets (Difference in % EER for All the Feature Sets is Highlighted by Oval). After [9].

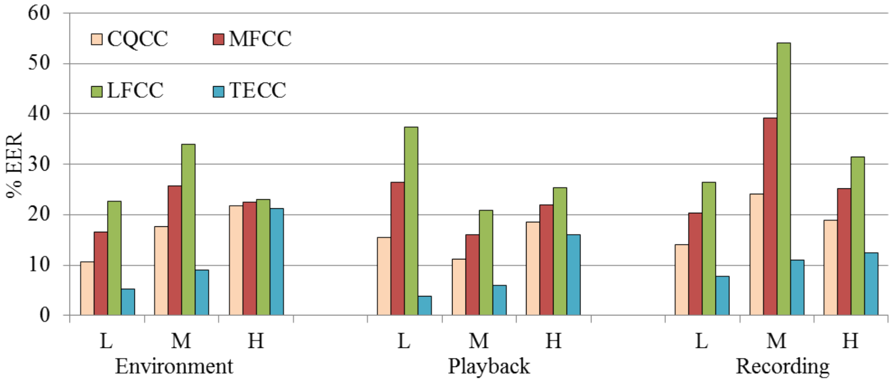


Figure 3.28: Different Levels of Threats, Namely, Low (L), Medium (M), and High (H) for CQCC, LFCC, MFCC, and TECC Feature Sets on All the Replay Configurations. After [9].

3.8.4 Results on ASVspoof 2019 Challenge Database

The organizers of ASVspoof 2019 challenge provided a baseline system for both Logical and Physical Access (LA and PA) tasks [7]. We observed in the study [9] that the spectral energy density obtained from the Teager energy-based approach has high energy across entire frequency regions (because of linearly-spaced Gabor filterbank) as compared to the spectral energy density obtained from the traditional spectrogram, and Moifies Group Delay (MGD) spectrum. The baseline system utilizes two feature sets, namely, CQCC and LFCC with 512 Gaussians for modeling genuine and corresponding spoof models in GMM. The ASVspoof 2019 challenge uses minimum Tandem- Detection Cost Function (t-DCF) as evaluation metric along with EER [7]. Due to computational load for the available hardware, less Gaussian mixtures were used for TECC feature extraction (i.e., 256 Gaussians for LA and only 64 Gaussians for PA task). From Table 3.10, it can be observed that TECC features outperform than the baseline systems. The results for the PA task is reported in Table 3.11. The training set of PA task, contains twice training files that were present in LA set, which in turn increases the computational load on the hardware and hence, we reduce Gaussian mixtures further. From Table 3.11, it can be observed that TECC feature set did not perform well for PA task, though it perform best on LA task.

Table 3.10: Comparison of TECC Feature Set with the Other Systems for LA Task of ASVspoof 2019 Challenge Database. After [9]

	Dev		**Eval**	
Feature Sets	EER	t-DCF	EER	t-DCF
CQCC	0.43	0.0123	9.57	0.2366
LFCC	2.71	0.0663	8.09	0.211
TECC	**0**	**0**	**7.51**	**0.1940**

Table 3.11: Comparison of TECC Feature Set with the Other Systems for PA Task of ASVspoof 2019 Challenge Database. After [9]

	Dev		**Eval**	
Feature Sets	EER	t-DCF	EER	t-DCF
CQCC	**9.87**	**0.1953**	11.04	0.2456
LFCC	11.96	0.2554	13.54	0.3017
TECC	24.7	0.62441	43.62	0.8085

In addition, the comparison of ASVspoof 2017, ASVspoof 2019, and real PA of ASVspoof 2019 challenge databases are shown in Table 3.12. On ASVspoof 2017 challenge database, an EER of 10.80 % and 11.41 % is obtained with TECC feature set on development and evaluation sets, respectively. The similar set of features did not perform well on the controlled acoustic environment, i.e., ASVspoof 2019 challenge database that results in 24.7 % and 43.62 % EER on development and evaluation sets, respectively. The absolute difference on development set for both replay databases is approximately 15 %, which is a huge difference for SSD task. The performance of the SSD system degrades in case of ASVspoof 2019 challenge database as this database is simulated, and has controlled acoustic conditions. This indicates that the same feature set do not work for different acoustical conditions, and hence, there is a need for more generalized features for SSD task. Furthermore, when the experiments were performed on real PA database of ASVspoof 2019, the EER is reduced from 43.62 % to 39.16 %. This indicate that uncontrolled acoustic environment indeed help to detect the replay signal from the natural speech.

Table 3.12: Results in % EER on ASVspoof 2017, ASVspoof 2019, and Real PA challenge database. After [9]

Feature	ASVspoof 2017		ASVspoof 2019		Real PA
Sets	Dev	Eval	Dev	Eval	Eval
CQCC	12.81	19.04	**9.7**	**11.04**	**15.71**
TECC	**10.80**	**11.41**	24.7	43.62	39.16

3.9 Chapter Summary

In this Chapter, we presented potential TEO-based features to capture reverberation during replay mechanism for SSD task. The genuine speech signal, $s[n]$, is inevitably convolved with the impulse response of the transmitting loudspeaker, $h_{mic}[n]$, before further convolution with the response of the channel under test, $h_1[n]$. We observed that for different acoustical environments, the Teager energy traces obtained are distinct. In particular, for a closed room (such as bedroom, office.) extra energy traces are observed because of echo impulses. In particular, suitability of TECC feature to capture reverberation characteristics depends upon acoustic environments for example, TECC may not perform well in outdoor environment where there is NO reverberation. Hence, these observations motivated us to extract features that are based on the energy traces, and thus, proposed TECC for replay SSD task.

The detailed theory of the TEO algorithm and feature extraction process was also presented. Results are shown for the SSD task on the ASVspoof 2015, ASVspoof 2017, BTAS 2016, and ASVspoof 2019 challenge databases. We observe that TECC feature set gave lower EER for all the different conditions of threat compared to the other system. For high-level threat and high quality devices used during playback, and recording, the EER are quite high. This needs further investigation to detect the high level replay configuration threat. In the next chapter, we will discuss ESA-based speech demodulation Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) features for SSD task.

CHAPTER 4

Temporal Modulation Features

4.1 Introduction

In this Chapter, we introduce the proposed feature sets used for SSD task, namely, Energy Separation Algorithm Instantaneous Frequency Cepstral Coefficients (ESA-IFCC), and Energy Separation Algorithm Instantaneous Amplitude Cepstral Coefficients (ESA-IACC). The proposed feature sets are based on Energy Separation Algorithm (ESA), and Teager Energy Operator (TEO). The basic concepts of the TEO and ESA are discussed in Section 4.2. The analysis of Instantaneous Frequency (IF) and Instantaneous Amplitude (IA) of a subband temporal modulation signal is also discussed along with the difference between natural and replay signal. The block diagram and feature extraction process along with importance of feature normalization is discussed in Section 4.3. The experiments are performed on ASVspoof 2017 and ASVspoof 2015 challenge corpora in Section 4.4.1 and Section 4.4.2, respectively. Finally, in Section 4.5, we summarize the Chapter.

4.2 Energy Separation Algorithm (ESA)

Traditional Fourier analysis requires the signal to be stationary, and the system under consideration should be linear. However, real-world signals, such as speech signals are non-stationary in nature (i.e., signals whose spectral contents vary with time) [170]. The non-stationary of speech signal is handled by short-term Fourier-transform (assuming speech is stationary within 20-25 ms range). To that effect, the representation of Amplitude Modulation- Frequency Modulation (AM-FM) signal is one of the approaches used for analysis of non-stationary signals. The AM-FM model is nonlinear in nature that describes a speech resonance as a signal with combination of AM and FM structure, i.e.,

$$R(t) = a(t)\cos\left[2\pi\left(f_c t + \int_0^t q(\tau)d\tau\right) + \theta\right], \tag{4.1}$$

where $R(t)$ is AM-FM speech signal, and f_c is the corresponding speech formant (center) frequency. The amplitude and frequency modulating signals of $R(t)$ are given by $a(t)$ and $q(t)$, respectively, and θ is a constant phase. The IF signal is defined as $f(t) = f_c + q(t)$, and speech signal, $s(t)$, is modeled as the sum, $s(t) = \sum_{k=1}^{K} R_k(t)$, where K is the number of AM-FM signals [188]. The basic problem in processing the AM-FM signal is demodulation, i.e., estimation of the information present in the AM and FM components. For monocomponent AM-FM signal, many successful demodulation approaches exist, such as Hilbert transform (HT) [183], Phase-Locked Loops (PLLs) [189], and Energy Separation Algorithm (ESA) [51].

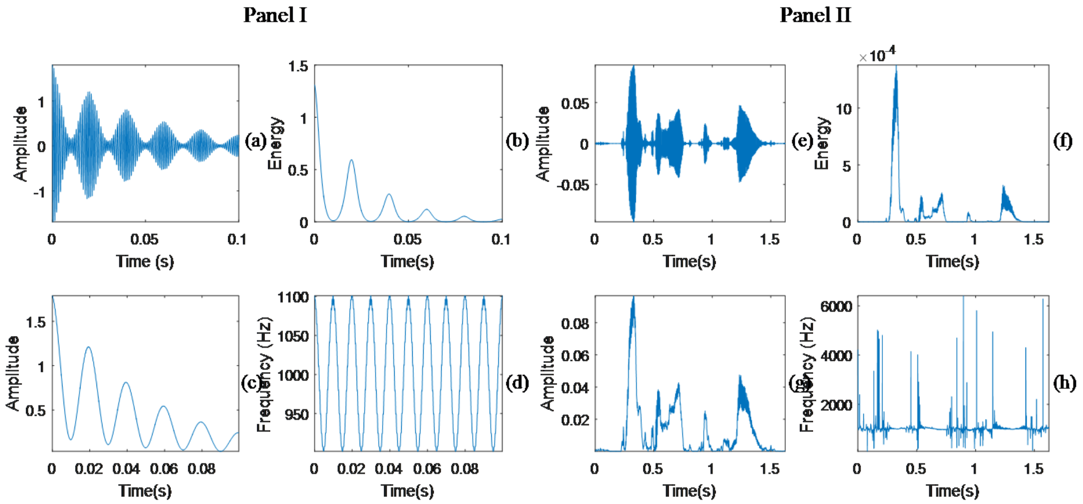


Figure 4.1: AM-FM Estimation Using ESA on a Synthetic (having closed form Expression) Signal (Panel I), and Natural Speech Signal (Panel II) for Utterance, *"Action Speaks Louder Than Words"*. (a) AM-FM Signal of $a[n]$=$0.998^n(1 + 0.8cos(pi/100)n)$, and $x[n]$=$a[n]cos((pi/5)n + sin(pi/50)n)$, (e) Filtered Narrowband Signal at f_c=1000 Hz, (c-g) Estimated IA, and (d-h) Estimated IF at f_c=1000 Hz for Synthetic Signal, and Speech Signal, (b-f) Shows Their Corresponding Teager Energies. After [12].

An example of synthetic (having closed-form expression) signal, and real natural speech signal is considered to show the estimated IA and IF components obtained from ESA demodulation approach in Figure 4.1. For synthetic case, the signal is considered with $a[n] = 0.998^n(1 + 0.8cos(pi/100)n)$, and $x[n] = a[n]cos((pi/5)n + sin(pi/50)n)$ shown in Panel I, and Panel II shows the voiced bandpass filtered signal with utterance, *"Action speaks louder than words"*, in Figure 4.1(a) and Figure 4.1(e). The speech signal is passed through a bandpass filter having a cut-off frequency of 1000 Hz. The corresponding IA and IF are shown in Figure 4.1(c-d) for synthetic signal whereas for speech signal, it is shown in Figure 4.1(g-h). The estimated IF in Figure 4.1(d-h) oscillates around its center frequency of 1000 Hz, the narrow spikes are usually caused either by amplitude valleys or by the onset of a new pitch pulse (i.e., T_0=1/F_0) [175].

Since speech signal has time-varying amplitude and frequency, it can be modeled as AM-FM signal [52]. Hence, TEO applied to an AM-FM speech signal can approximately estimate the squared product of IA ($a_i[n]$), and IF ($\Omega_i[n]$) for the i^{th} subband filtered signal, i.e.,

$$\Psi_d \left\{ a_i[n] cos \left(\sum_0^n \Omega_i[m]dm + \theta \right) \right\} \approx a_i^2[n]\Omega_i^2[n]. \tag{4.2}$$

The ESA is applied on a single speech resonance. However, speech is a combination of several resonances and hence, these resonances needs to be separated using bandpass subband filtering [190]. Out of several versions of discrete-time ESA (DESA) algorithm, DESA-2 (symmetric approximation) has the advantage of less computations per sample than the DESA-1 (asymmetric approximation) [57]. Hence, we used DESA-2 algorithm to estimate '$a_i[n]$' and '$\Omega_i[n]$' of the narrowband filtered signal, ESA was developed and is given by (Refer Appendix B for more details of ESA) [51,57,175]:

$$a_i[n] \approx \frac{2\Psi_d\{x_i[n]\}}{\sqrt{\Psi_d\{x_i[n+1] - x_i[n-1]\}}}, \tag{4.3}$$

$$\Omega_i[n] \approx arcsin\sqrt{\frac{\Psi_d\{x_i[n+1] - x_i[n-1]\}}{4\Psi_d\{x_i[n]\}}}. \tag{4.4}$$

In this chapter, we focus on the ESA approach to estimate the IA and IF components of a speech signal.

The physical significance in terms of temporal modulations at different time scale is analyzed in Figure 4.2. The time-domain subband filtered signal around 1^{st} formant frequency is shown in Figure 4.2(a) for natural (Panel I), and replay (Panel II). The slow temporal modulations of a speech signal roughly correlates with the different *syllabic* segments. For natural speech, slow temporal modulations results in smooth amplitude envelope as shown in Figure 4.2(b) (in Panel I). The higher peaks in the fast temporal modulations (which are also known as Temporal Fine Structure (TFS)) as shown in Figure 4.2(c) represents the *harmonic*

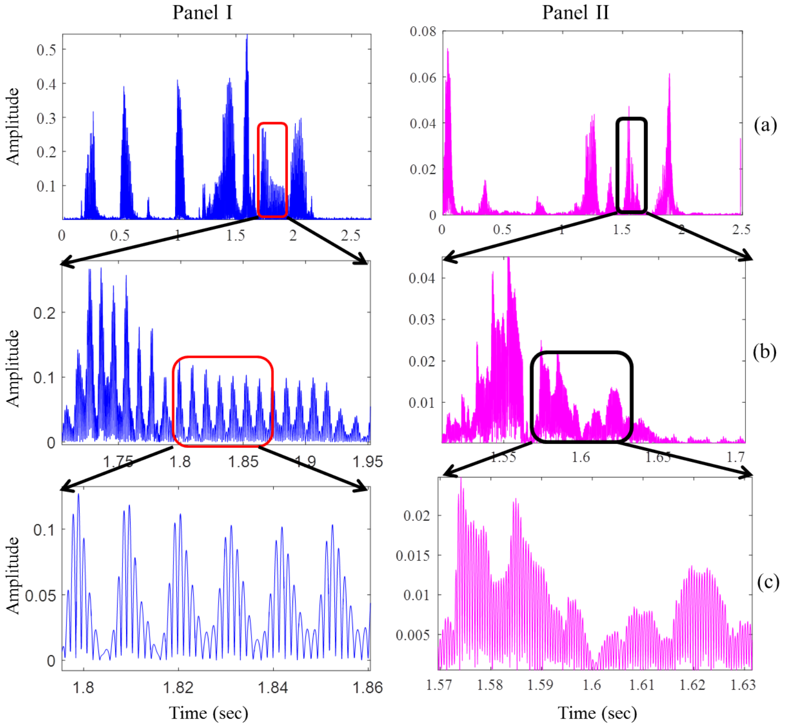


Figure 4.2: (a) Time-Domain Subband Filtered Speech Signal around first formant frequency, Whose Temporal Modulations are Depicted at Different Time Scales, (b) Modulations Due to the Inter-Harmonic Interactions, and (c) Fast Temporal Modulations. Panel I, and Panel II is for Natural and corresponding Replay Spoof Signal, respectively. After [12,26].

structure of the speech signal [26]. However, this observation is missing for the replay speech (Panel II) of Figure 4.2. The slow temporal modulations for replay speech are having distorted amplitude envelope (Panel II) of Figure 4.2(b). The fast temporal modulations do not represents the harmonic structure in Figure 4.2(c) of Panel II. This observation of natural and replay speech signal motivated us to analyze more on temporal modulations for replay SSD task.

The IA and IF components estimated from the subband filtered signal is shown in Figure 4.3 (d and e). The IA component is related to the slow temporal modulations of a speech signal Figure 4.3(b). On the other hand, the IF components shows the fluctuations driving around the cutoff frequency, $F_c = 494$ Hz (as shown in Figure 4.3(e)), and it is because of fast temporal modulations shown in Figure 4.3(c)) that represents the harmonic structure of the speech signal [26].

The comparison of the natural and replay speech signal in terms of IF components is observed (as shown in Figure 4.4). The segment of a speech signal with bandpass filtered around center frequency of 500 Hz for natural (Panel I), and replay (Panel II) is shown in Figure 4.4. The highlighted dotted rectangular box in Figure 4.4(a) shows the difference in time-domain waveform for the voiced re-

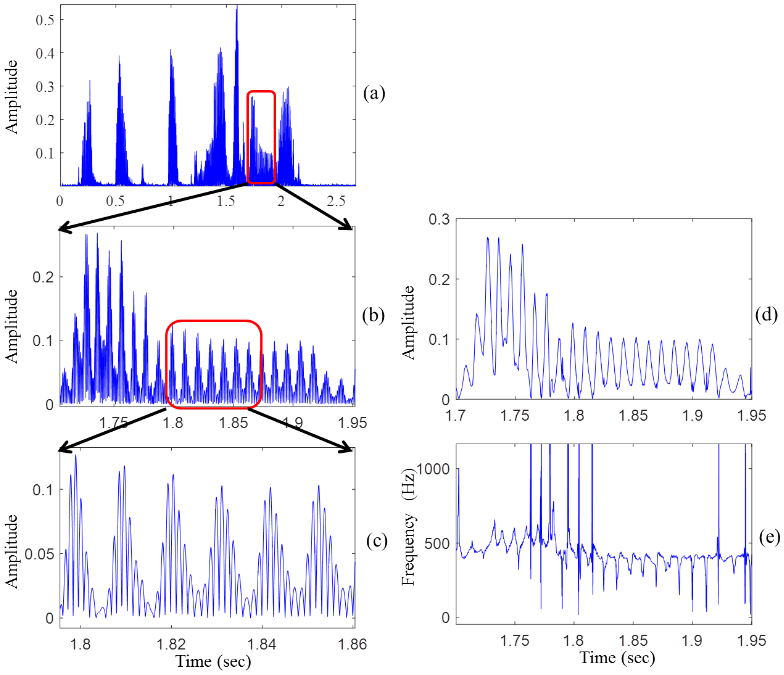


Figure 4.3: (a) Time-Domain Subband Filtered Signal (Using Gabor Filterbank) at Cut-Off Frequency, F_c = 494 Hz, Whose Temporal Modulations are Depicted to the Right at Different Time Scales, (b) Modulations Due to the Inter-Harmonic Interactions, (c) Fast Temporal Modulations, (d) Amplitude Envelope of the Response of (b) is Clearly Observed, and (e) Fast Temporal Modulations are Due to the Frequency Components [12,26].

gion of a natural and a replay signal. The voiced region in Panel I from 2.82 s to 2.9 s has the time-varying amplitude which is normal for natural speech, whereas for replay speech (Panel II), the same voiced region is changed to the pattern of sinc function-like. The replay speech signal consists of repeating sinc function-like pattern observed from 1.83 s to 1.9 s (as shown in Panel II of Figure 4.4(a)). The IF fluctuations for natural and replay around center frequency of 500 Hz is shown in (Panel I and Panel II) of Figure 4.4(b). It can be observed that from Panel I that the IF has fluctuation exactly around the center frequency of 500 Hz, whereas this is not the case for the corresponding IF fluctuations obtained from the replay speech. In particular, the IF fluctuates below the center frequency, and the damping of fluctuations starts exactly from the same time duration, i.e., 1.83 s to 1.9 s, as the pattern starts with sinc function-like pattern for replay speech.

Figure 4.5(a) shows the plot of a genuine speech utterance and Figure 4.5(b-c) shows the respective IA and IF of a narrowband filtered signal around 1500 Hz for a speech signal shown in Fig. 4.5(a). The output is possibly decomposed into its corresponding AE and IF from the filterbank of various narrowband component. The IF in Figure 4.5(c) is centered around 1500 Hz, and shows spurious fluctua-

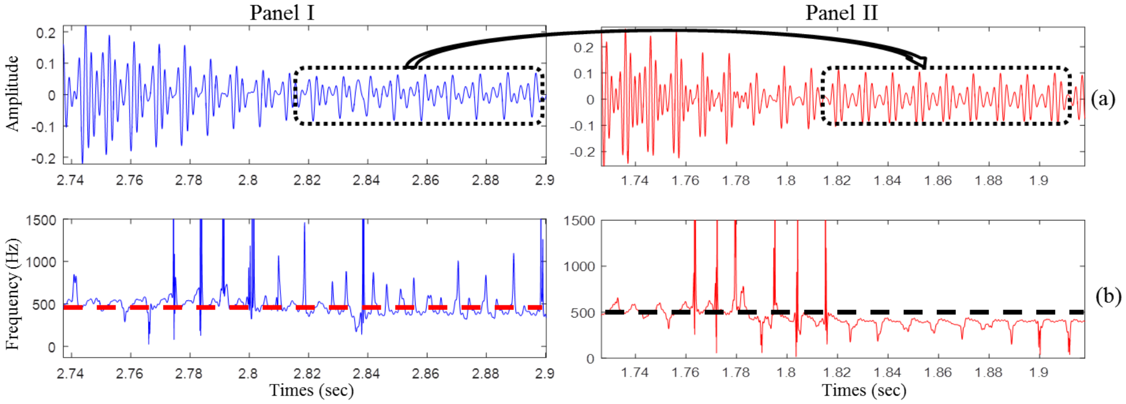


Figure 4.4: (a) Subband filtered segment of a speech signal around a center frequency 500 Hz of natural speech signal (Panel I), and corresponding subband filtered replay signal (Panel II) segment, and (b) shows the corresponding IF fluctuations of natural and replay speech segment. Highlighted region shows the change in the sinc function-like pattern for replay speech signal (Panel II(a)), and the damping of IF components from the center frequency for the same speech segment. After [12].

tions on both side that makes it difficult to analyze and interpret the vocal tract system characteristics [167]. There could be two reasons for IF that has spurious fluctuations in a speech signal and they are:

- When the amplitude approaches to zero, then the large fluctuations are observed for IF. For the region 4.4 s to 4.5 s in Fig. 4.5(c), i.e., unvoiced regions have more changes because of narrowband components have low energy in that region as shown in Fig. 4.5(d).

- On the other hand, the region from 4.2 s to 4.25 s and 4.3 s to 4.4 s as in Fig. 4.5(c), which is voiced regions have the changes in IF connected to impulse-like nature during speech production [167].

This is because the IF estimated from the speech signal contains the information of both vocal tract system, and speech excitation source. As a result, IF shows impulse-like discontinuities at the instants of glottal closure [191]. As a result, at the instants of Glottal Closure Instants (GCI), discontinuity of impulses are observed in IF. The impulse response of vocal tract system during production of a speech signal originated GCI successively to yield a speech signal. The phase discontinuity occur due to superposition of impulse response that gives us the large amplitude peaks in IF at locations of GCI.

The key advantages of ESA algorithm is that it does not require the complex task of phase unwrapping (as required in HT-based approach) and in addition, *only* five samples are required to estimate $a_i[n]$ and $\Omega_i[n]$ and thus, avoid the need

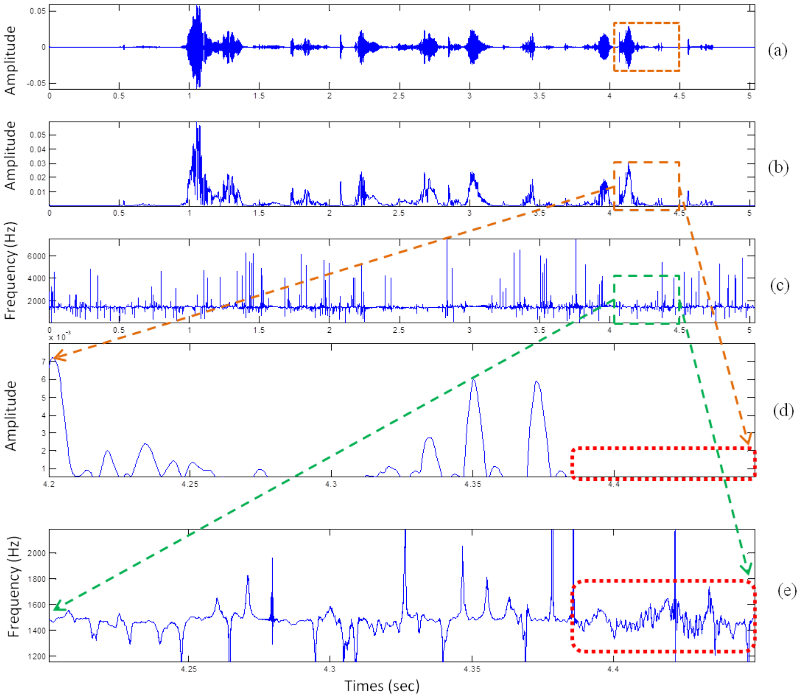


Figure 4.5: AM-FM Decomposition (a) Speech Signal, (b) IA, (c) IF of Filtered Narrowband Signal at f_c=1500 Hz, and Fig. 4.5(d)-(e) Amplitude Envelope, and Instantaneous Frequency, respectively, of Speech Segment Shown With Dotted Box in Fig. 4.5(a). After [13].

for segmental block-based processing of speech, which has its own disadvantages, such as the positioning of the window, length of analysis window, and accuracy of IF estimation decreases primarily due to block-based processing [191].

4.3 Details of Proposed Feature Set using ESA

Recent studies observed that the spectral information present in higher frequency regions is more distorted for replay speech [135]. Hence, to emphasize these high frequency regions, pre-emphasis filter is used having a system function, $H(z) = 1 - az^{-1}$, with a typical filter coefficients of $a = 0.97$ [169]. This pre-emphasized speech signal is passed through a Gabor filterbank to obtain narrowband filtered signals. We used Gabor filter since it is *compact* and *smooth*. The Gabor filter has *optimal* joint time-frequency resolution (since Fourier transform (FT) of a Gaussian is a Gaussian, and it is infinitely differentiable function, i.e., $g(t) \epsilon C^{\infty}$) [170]. The impulse response, $g(t)$, of a Gabor filter is given by [52] :

$$g(t) = exp(-b^2t^2)cos(\omega_c t), \tag{4.5}$$

where ω_c is the center frequency (in Hz) of the subband filter chosen as per the frequency scales of Equivalent Rectangular Bandwidth (ERB), Mel, and linear. The parameter b controls the bandwidth of a filter. The Gabor filters have the linear phase response characteristics and hence, it maintains the same pattern (shape) of the filtered speech signal (within the passband of subband filter) [171]. Motivated by the studies of auditory perception mechanism for humans, the center frequencies of ERB, and Mel scales have narrow and wider bandwidth in the lower and higher frequency regions, respectively [172,173]. In case of linear frequency scale, all the subband filters have almost equal bandwidth and hence, have good resolution in the lower, and higher frequency regions that make more reliable to extract high resolution spectral information.

4.3.1 Frequency Scale

4.3.1.1 ERB Frequency Scale

This frequency scale is used to quantify the bandwidth of asymmetrical filters like the auditory filter. The study reported in [172] observed that the auditory filter bandwidths are given as:

$$ERB(\omega) = 6.23(\omega/1000)^2 + 93.39(\omega/1000) + 28.52. \tag{4.6}$$

4.3.1.2 Mel Frequency Scale

Stevens *et al.* in 1937 proposed Mel scale, which is perceived pitches (subjective) similar to the perception prediction of the human ear [173]. Spectral resolution of Mel scale becomes lower as the frequency increases. The mathematical expression of Mel scale is given as:

$$Mel(\omega) = 2595log_{10}(1 + \omega/700). \tag{4.7}$$

4.3.1.3 Linear Frequency Scale

In a linear frequency scale, bandwidth is equally distributed throughout all the frequency ranges that makes it more reliable to extract the proposed features. It effectively capture the information in the lower as well as higher frequency regions. The mathematical expression of linear scale is given as:

$$Lin(\omega) = \omega. \tag{4.8}$$

The placing of center frequencies according to the frequency scale is shown in Figure 4.6. We can see that with ERB and Mel scale, only 20 subband filters are covered with approximately 1000 Hz frequency, whereas for linear frequency scale to cover 20 subband filters, approximately 4000 Hz frequency is used. The linear frequency scale has good frequency resolution in both lower and higher frequency regions because of its equal distribution of bandwidth across the entire frequency regions.

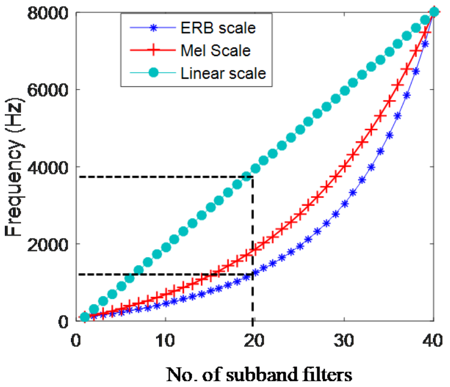


Figure 4.6: Placing of 40 Number Center Frequencies w.r.t ERB, Mel, and Linear Frequency Scale Upto f_{max}=8000 Hz. After [12].

4.3.2 Feature Normalization

Most of the ASV systems perform well under clean environmental conditions. The performance of ASV system is affected by the acoustic mismatch between acoustic environments for training and testing [192, 193]. The speech signals are recorded in various environmental conditions that includes the mismatch problem due to distortion involve in speech signal [5]. The Cepstral Mean Normalization (CMN) is a simple technique to reduce this channel mismatch problem [174], it is also known as Cepstral Mean Subtraction (CMS) [194]. The basic principle behind CMN technique is the nature of the cepstrum under the convolution distortions operations [193]. The impulse response of a channel, $h[n]$, is assumed to be Linear Time-Invariant (LTI) system. Let us denote clean speech signal as, $s[n]$, and channel distortion as, $h[n]$, the corrupted speech signal is the convolved speech signal of clean, and channel distorted signal denoted by $y[n]$:

$$y[n] = s[n] * h[n]. \tag{4.9}$$

In cepstral-domain, the convolution operation gets converted to addition operation, and is given by Eq. (4.10) [193, 194]:

$$c_y = c_s + c_h, \tag{4.10}$$

where c_y, c_s, and c_h, represent the cepstral representation of $y(n), s(n)$, and $h(n)$. Now, calculate the mean of Eq. (4.10), i.e.,

$$E[c_y] = E[c_s] + E[c_h], \tag{4.11}$$

where $E[\cdot]$ is the expected value and it is the time average across the long recorded feature frames. Since it is assumed that channel, $h[n]$, does not change over the duration of an utterance, so average of c_h becomes the cepstrum of the channel, $c_{(h)}$. If the variation and distortion of sounds in $s_{(n)}$ is such that the average spectrum over the utterance is relatively flat, then average value of clean speech E$[c_s]$ ≈ 0. However, the cepstral mean of corrupted speech signal, $y[n]$, has just become the cepstrum of the channel, i.e., $E[c_y] = c_h$. Now, remove the channel distortion by subtracting cepstral mean, $E[c_y]$, from the cepstral of corrupted speech signal, c_y [193], i.e.,

$$c_y = c_y - E[c_y]. \tag{4.12}$$

$$c_y = c_s + c_h - c_h, \tag{4.13}$$

$$\therefore c_y = c_s. \tag{4.14}$$

Thus, we can reduce the channel distortion by subtracting the cepstral mean, $E[c_y]$ from the cepstral of a corrupted speech signal. From Eq. (4.14), the cepstral of corrupted speech signal is free from the channel distortion.

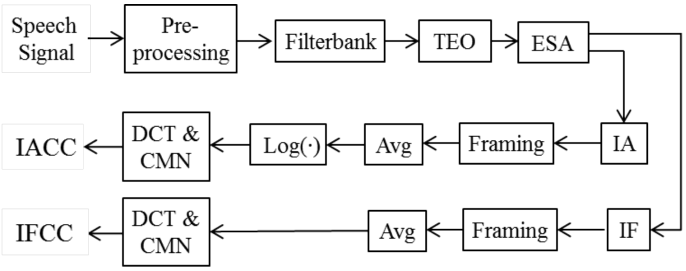


Figure 4.7: Schematic Block Diagram of the Proposed Feature Sets, namely, Instantaneous Amplitude and Frequency Cepstral Coefficients (IACC and IFCC). After [27].

The proposed feature set is extracted as per the block diagram as shown in Figure 4.7. The ESA method is applied on the Teager energy profiles, and estimates the IA and IF components for each filtered subband speech signal. Furthermore, these IAs and IFs profiles are passed through the frame-blocking, and averaged with a short window of length 20 ms and with a shift of 10 ms followed by logarithm operation to compress the data (in a way similar to the human auditory system) [195]. However, in our earlier studies in [27], we observed that for estimating IA components with logarithm operation provides lower Equal Error

Rate (% EER), whereas IF estimated by logarithm operation did not give better results and hence, IF components were estimated without using logarithm operation. To obtain a low-dimensional representation, Discrete Cosine Transform (DCT) is applied and retained first few DCT coefficients, namely, Instantaneous Amplitude Cepstral Coefficients, and Instantaneous Frequency Cepstral Coefficients (i.e., IACCs and IFCCs). The performance of ASV system is known to be affected by the mismatch environmental conditions between the training and testing [192]. The replay database was recorded in various environmental conditions that include the mismatch problem due to distortion involved in speech signal [5]. The Cepstral Mean Normalization (CMN) (also known as Cepstral Mean Subtraction (CMS)) is used to reduce the channel mismatch/distortion conditions [174]. Hence, to reduce the channel mismatch conditions, these IACCs and IFCCs were further processed with CMN technique, and appended along with their Δ and $\Delta\Delta$ features in order to obtain higher-dimensional feature vector.

4.4 Experimental Results

In this Section, the results are shown for several experimental evaluation factors on ASVspoof 2017 v 2.0 challenge database. The extraction of the proposed feature sets is mainly affected by the parameters of the filterbank, namely, the shape of the subband filter, the choice of frequency scale, and the number of subband filtered signals. In addition, we also performed experiments on ASVspoof 2015 challenge database.

4.4.1 Results on ASVspoof 2017 Challenge

4.4.1.1 Results with Butterworth and Gabor Filterbank

In our earlier studies, we used Butterworth filterbank to extract the feature sets [13, 28]. We compared the results with linearly-spaced Butterworth, and Gabor filterbanks. The spectral energy density obtained with 40 number of subband filtered signals are shown in Figure 4.8. The time-domain speech signal for the utterance, *"Action speaks louder than words,"* is shown in Figure 4.8 (a), whereas Figure 4.8 (b), Figure 4.8 (c), and Figure 4.8 (d) shows the spectral energy density obtained from the traditional spectrogram along with Butterworth and Gabor filterbanks, respectively. In particular, proposed approach brings out more dominant regions of spectral energy densities than the traditional spectrogram (indicated via directional arrows in Figure 4.8(b) to Figure 4.8(d)). The spectral energies obtained from the Gabor filterbank preserves good resolution in both lower and the higher frequency regions (highlighted with the solid box and oval) than the spectral energies obtained from the Butterworth filterbank. This lower frequency information provides the lower formant information (i.e., F_1 and F_2), and the higher formants (i.e., F_3 and F_4) are present in the higher frequency regions [196]. This spectral energies obtained from the Gabor filterbank helps to capture the information present in natural and replay speech signal. The effect of Gabor filterbank is to smooth out the spikes, and the abrupt jumps (if any) of the original estimates, where high frequency components are preserved [197].

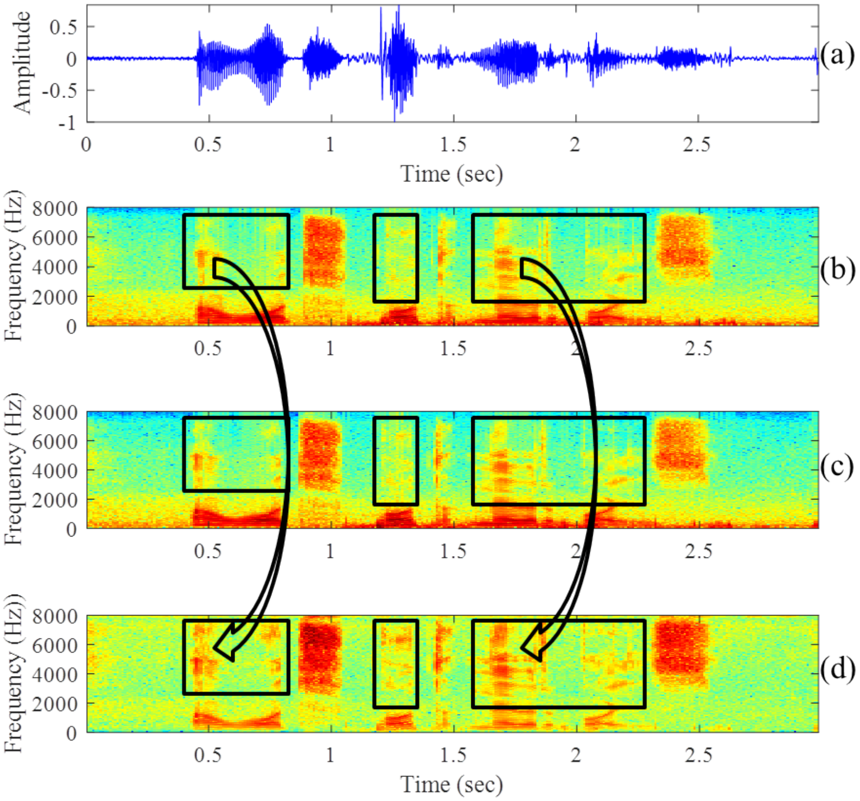


Figure 4.8: Comparison of Spectral Energy Density Between Butterworth and Gabor Filterbank. (a) Time-Domain Speech Signal, Spectral Energy Density of (b) Traditional STFT Spectrogram, (c) Butterworth Filterbank, and (d) Gabor Filterbank Obtained with 40 Subband Filtered Signals. Highlighted Regions Shown via Rectangular Boxes Indicates the Spectral Energy Differences Obtained From Different Techniques. The Direction of Arrows From Figure 4.8(b) to Figure 4.8(d) Indicates That the Proposed Approach is Able to Bring Dominant Spectral Energy Density than the Relatively Weaker Density Regions in to the Traditional Spectrogram. After [12].

Moreover, impulse response of Butterworth filter is Infinite Impulse Response (IIR) in nature and hence, has the nonlinear phase response while, Gabor filter has the linear phase response characteristics [171, 198]. The nonlinear phase in frequency-domain of this acoustic system is added to the phase of natural speech and hence, the results with Butterworth filterbank did not give better results [60]. The results of proposed feature set obtained with linearly-spaced Butterworth, and Gabor filterbanks having *120*-D (Static (S)+Δ+ΔΔ) feature vector are shown in the bar graph representation in Figure 4.9. It can be observed that the % EER obtained on development set with Gabor filterbank for the proposed feature sets is lower. In particular, ESA-IACC feature set gave an EER of 7.99 % which is much lower than the results obtained from the Butterworth filterbank (17.20 %). Thus, the choice of a filterbank should be considered as an important factor for better estimation of IA and IF components. Hence, further set of experiments in this Chapter are performed using Gabor filterbank.

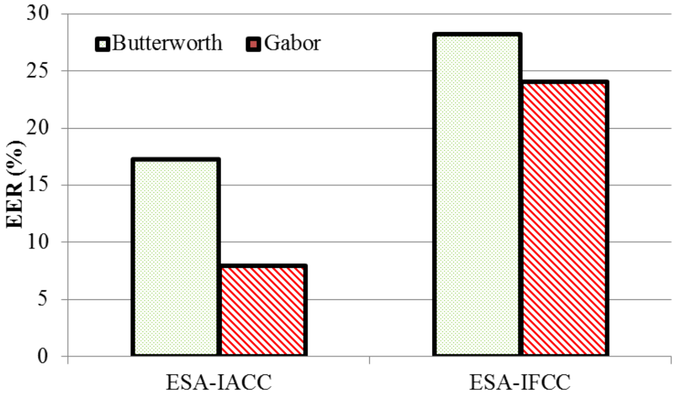


Figure 4.9: Bar Graph Representation of Proposed Feature Sets with Butterworth, and Gabor Filterbank. After [12].

In addition, we also performed an experiment without integrating Gabor filterbank to investigate the importance of filterbank for the proposed approach (i.e., filterbank block in Figure 4.7). Table 4.1 shows the results (with and without filterbank) on development and evaluation sets for ESA-IACC and ESA-IFCC feature sets. It can be observed that the EER obtained for both the feature sets gave high EER for both development and evaluation set.

Table 4.1: Results on ASVspoof 2017 Database Obtained With and Without Applying Filterbank Along with CMN on Proposed Feature Sets. After [12]

	Without Filterbank		With Filterbank	
Feature Set	Dev	Eval	Dev	Eval
ESA-IACC	23.37	40.47	**7.99**	**13.45**
ESA-IFCC	16.57	39.23	**11.84**	**12.93**

4.4.1.2 Effect of Frequency Scales

In this sub-Section, we investigate the effect of different frequency scales used during estimation of IA and IF-based feature sets in the Gabor filterbank. Here, we have studied three types of frequency scales, namely, ERB, Mel, and linear scale. The results obtained with different frequency scale are shown in Figure 4.10. It can be observed that the EER is reduced with linear scale for both the proposed feature sets (highlighted by oval dotted circle in Figure 4.10). For ESA-IACC feature set, the EER obtained with ERB scale is 26.58 %, and it is reduced to 7.99 % with linear scale. Similarly, for ESA-IFCC feature set, the EER obtained with ERB scale is 36.17 %, and it is reduced to 24.07 % with the linear scale. Hence, the remaining experiments in this Chapter for estimating IA and IF are carried out with the linear frequency scale.

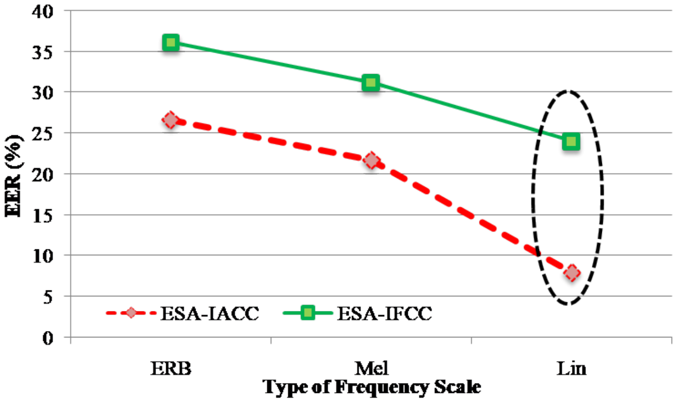


Figure 4.10: Results of Proposed Feature Sets w.r.t. Different Frequency Scales, namely, ERB, Mel, and Linear (Highlighted Oval Shows the Decrease in % EER). After [12].

4.4.1.3 Effect of Number of Subband Filters

The human auditory system carries several thousands of subband filters which results in a dense filterbank in frequency-domain [53]. Thus, in this Section, we experimented on effect of increasing the number of subband filters. The estimation of IA and IF components using ESA can be obtained more accurately when the speech signal is filtered to a narrowband signal. Hence, to obtain the narrowband filtered signal, we compute the results by increasing the number of subband filtered signals. Since the proposed feature sets are extracted with linearly-spaced subband filters, the frequency resolution is explicitly related to the number of subband filters. The subband filtered signals should be chosen such that it covers the entire spectrum information. When the number of subband filters used in filterbank is less, it might loose the information and if the subband filters are more, then it may get overlap with the adjacent subband filters.

1. *Increasing Subband Filtered Signals for IA Features*: This Section presents the experiments performed by varying the number of subband filtered signals for IA feature set. Initially, for estimating the IA components, we kept 40 subband filters in Gabor filterbank with 200 Hz bandwidth. We further increased the number of subband filters up to 140 in filterbank. It can be observed that by increasing the number of subband filters, the EER keeps on increasing as shown Figure 4.11. Relatively least EER of 7.99 % is obtained with 40 subband filters in filterbank. The amplitude spectrum is smooth compared to the variations/fluctuations estimated from IF components. By increasing the number of subband filters in filterbank, IA components might get overlapped with adjacent subband filters and hence, the computation becomes inaccurate. In this case, only 40 subband filters are required to cover the entire spectrum information (i.e., available bandwidth) that can be *discriminative* for the replay SSD task.

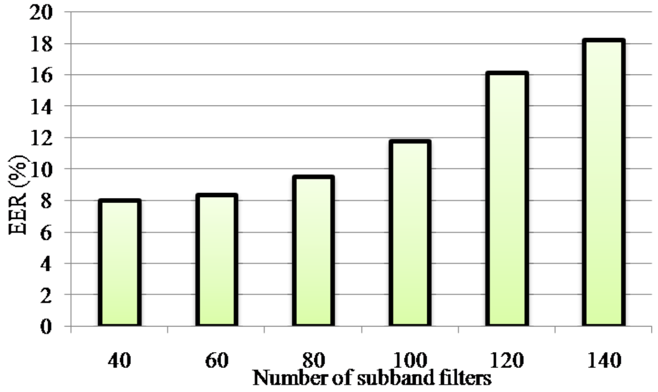


Figure 4.11: Results of ESA-IACC Feature Sets with Varying the Subband Filters. After [12].

2. *Increasing Subband Filtered Signals for IF Features*: On the other hand, for estimation of IF components, we have analyzed the effect of increasing the number of subband filters on 200 Hz and 400 Hz BW as shown in Figure 4.12. The estimation of IF estimation is exactly opposite to that of IA estimation when it comes to the choice of number of subband filters in filterbank [53]. From Figure 4.12, we can see that as the number of subband filters is increased, the EER keep on decreasing and later, we get a near-constant EER (i.e., no further improvements in SSD performance). The least EER of 11.4 % is obtained with 120 number of subband filters in a filterbank with 200

Hz bandwidth. Hence, for further set of experiments, we have selected 40 and 120 number of subband filtered signals in a filterbank for extraction of ESA-IACC and ESA-IFCC feature sets, respectively.

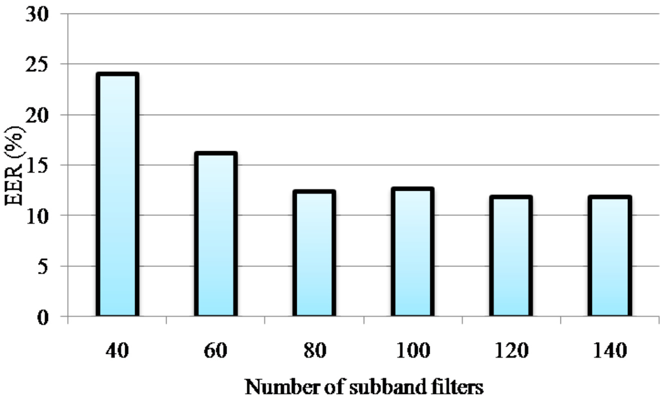


Figure 4.12: Results of ESA-IFCC Feature Sets with Varying the Subband Filters. After [12].

4.4.1.4 Results on Score-Level Fusion

On the development set, it is observed that instead of using Butterworth filterbank, the Gabor filterbank gives lower % EER. In addition, we also observed that the linearly-spaced center frequency in Gabor filterbank gives lower % EER compared to the ERB and Mel frequency scales. Furthermore, we also observed that the estimation of IA and IF components gave the least % EER when the number of subband filtered signals in a filterbank are kept to 40 and 120, respectively. Based on the results obtained on the development set, similar parameters are chosen for evaluation set. Table 4.4 shows the results in % EER on the evaluation set of proposed feature sets. We have compared our proposed feature sets with CQCC, MFCC, and LFCC. On the ASVspoof 2017 version 2.0 database, the baseline is modified from the earlier baseline in the form of having the log-energy coefficients, and CMVN method. With the enhanced baseline, the results are as shown in Table 4.4. However, we have considered the CQCC feature set with CMN as our first baseline (abbreviated as CQCC_CMN) as these features are in cepstral-domain so that we can compare with the other feature sets. In addition, we have also considered LFCC as our second baseline, since proposed feature sets are extracted with the linear frequency scale, and the algorithm used for proposed feature sets are based on the linear frequency scale. We also compared results with MFCC feature set as it is one of the state-of-the-art feature set used for various applications in the speech literature. Furthermore, we also performed the experiment by applying the CMVN method for proposed feature sets. In general, we found CMN is better than the CMVN technique. Thus, features extracted using CMN (which is an approximate highpass filter) are effective for replay SSD task. Since in replay spoof speech the higher frequency regions are more discriminative compared to the lower frequency regions, CMN helps to apply highpass filtering in speech and thus, boost all the more the discriminative high frequency regions. In addition, replay signal being affected due to bandpass nature of acoustic environment, the highpass filter may help to capture that effect. The results obtained with CMVN (which is equivalent to adaptive gain control) for proposed feature set gave high EER as shown in Table 4.2 and hence, we extract the proposed feature sets by applying CMN method as it gave relatively low EER for ESA-IACC, and ESA-IFCC feature sets.

Table 4.2: Results on Development and Evaluation Set Using CMN and CMVN Technique Using 40 Subband Filtered Signals. After [12]

	Dev		Eval	
Feature Set	CMN	CMVN	CMN	CMVN
ESA-IACC	**7.99**	12.48	**13.45**	34.04
ESA-IFCC	**24.32**	27.29	**19.87**	29.32

Table 4.3: The % EER for Score-Level Fusion at Various Fusion Factors on Evaluation Set. After [12]

Fusion Factor	Feature sets used for score-level fusion						
α	CQCC+ESA-IACC	CQCC+ESA-IFCC	MFCC+ESA-IACC	MFCC+ESA-IFCC	LFCC+ESA-IACC	LFCC+ESA-IFCC	ESA-IACC+ESA-IFCC
0	19.04	19.04	26.90	26.90	15.73	15.73	12.93
0.1	15.71	15.08	20.88	19.51	14.68	14.32	12.15
0.2	13.60	13.21	17.24	18.06	14.03	13.51	11.38
0.3	12.62	11.74	14.94	16.97	13.62	13.07	11.19
0.4	12.11	10.89	13.65	15.76	13.36	12.75	11.10
0.5	11.99	10.33	12.95	14.81	13.12	12.39	**11.09**
0.6	**11.93**	**10.12**	**12.65**	13.90	13.08	12.10	11.30
0.7	12.12	10.30	12.78	13.26	**13.01**	**11.83**	11.62
0.8	12.45	10.90	12.95	12.85	13.13	12.08	11.87
0.9	12.89	11.71	13.12	**12.81**	13.23	12.37	12.19
1	13.45	12.93	13.45	12.93	13.45	12.93	13.45

To explore the possible complementary information, we have used score-level fusion (as per Eq. (6.7)). For development set, the relatively best % EER was obtained with α =0.5 for CQCC and ESA-IACC. On the other hand, for CQCC and ESA-IFCC feature sets, the best % EER was obtained with α =0.7 (as shown in Figure 4.13). Similar pattern of reduced EER was observed on the evaluation set, i.e., when CQCC is used with proposed feature sets using score-level fusion as shown in Table 4.3. For evaluation set, the relatively least % EER was obtained with fusion factor of α =0.5. The least % EER obtained on development set is 7.03 % with fusion of CQCC and ESA-IACC, while for evaluation set, the least EER of 10.12 % is obtained with fusion of CQCC and ESA-IFCC feature sets.

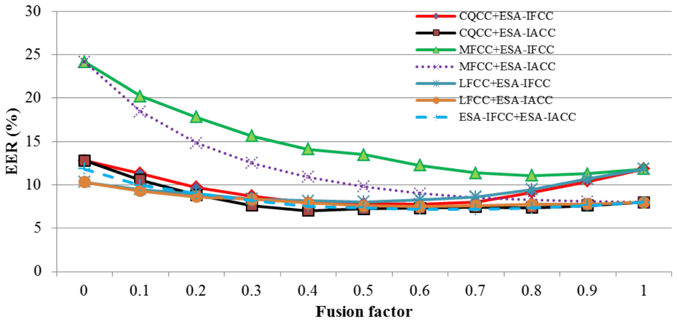


Figure 4.13: The % EER for Score-Level Fusion at Various Fusion Factors on Development Set. After [12].

Table 4.4: Final Results on ASVspoof 2017 Challenge Version 2.0 Database (in % EER) on Development and Evaluation Set. The Parenthesis Bracket Shows the Weight of Fusion Parameter α. After [12]

Feature Set	Dev	Eval
CQCC (Post-eval BL)	9.06	13.74
A: CQCC-CMN	12.81	19.04
LFCC (Our Baseline2)	16.76	13.90
MFCC	24.19	26.90
ESA-IACC (CMVN)	13.45	34.04
ESA-IFCC (CMVN)	19.87	29.32
B: ESA-IACC	7.99	13.45
C:ESA-IFCC	11.84	12.93
A+B	**7.03** (α =0.5)	12.11 (α =0.5)
A+C	7.74 (α =0.7)	**10.12** (α =0.7)
B+C	7.13 (α =0.6)	11.09 (α =0.6)

BL: Baseline

In addition, in order to capture the possible complementary information present in IA and IF components-based features, we have used their score-level fusion (as shown in Table 4.4). We observed that the performance obtained from the fusion of IA and IF components gave the reduced % EER over the individual counterparts. These reductions in % EER indicate that both IA and IF-based feature sets are complementary in nature. The least % EER of 7.13 % (development set, with a fusion factor of α =0.6), and 11.09 % (evaluation set, with a fusion factor of α =0.6) is obtained by the fusion of ESA-based features using GMM classifier.

4.4.1.5 Analysis of Replay Configurations

The replay speech signals are recorded in different acoustic environments using different recording, and playback devices [5]. These intermediate devices are further classified into three different levels of threats, namely, low, medium, and high. Figure 4.14 shows the performance of CQCC, LFCC, MFCC, ESA-IFCC, and ESA-IACC feature sets for all the levels of threats. In particular, for acoustic environment, recording, and playback device, it can be observed that the proposed feature set has lower % EER in each case as compared to the other feature sets. The high-level threats pose more challenging for recording and playback devices since very high quality devices are used to record and playback the replay signal. Hence, these replay signals are very much similar to their natural counterparts and thus, degrades the performance resulting in higher % EER for SSD system.

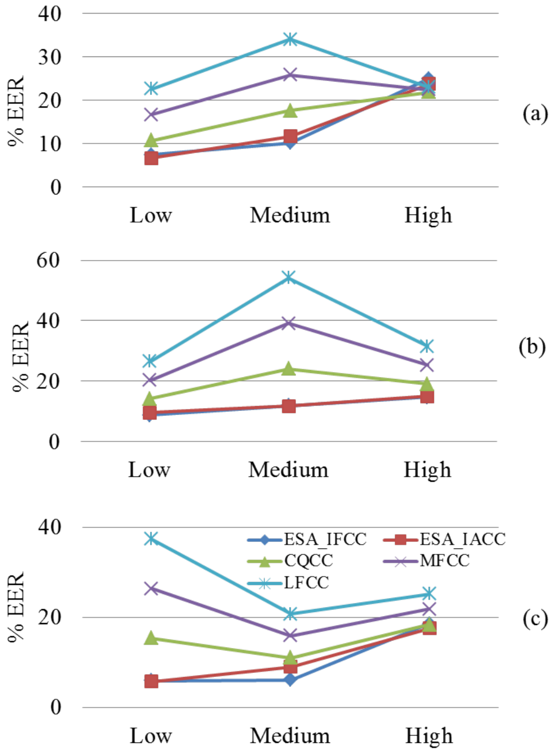


Figure 4.14: Results in % EER for Different Levels of Threats on Replay Configurations, Namely, (a) Acoustic Environment, (b) Recording Device, and (c) Playback Device with CQCC, MFCC, LFCC, ESA-IFCC, and ESA-IACC Feature Sets. After [12].

The acoustic environment listed in [6] are the actual space in which the original speech data is replayed and re-recorded. The ASVspoof 2017 challenge version 2.0 database have 26 different environments denoted from E01-E26. Different environments have the variations with the levels of additive ambient, convolutive, and reverberation noise. The Figure 4.15(a) shows the detailed % EER for all the different environmental conditions with all the feature sets on evaluation dataset. It can be observed that for CQCC, MFCC, and LFCC feature sets, the % EER for all the environmental conditions are high compared to the TECC feature set. Hence, TECC feature set (red line) shows the lower % EER for all the different environmental conditions. Similar to different acoustic environments, there

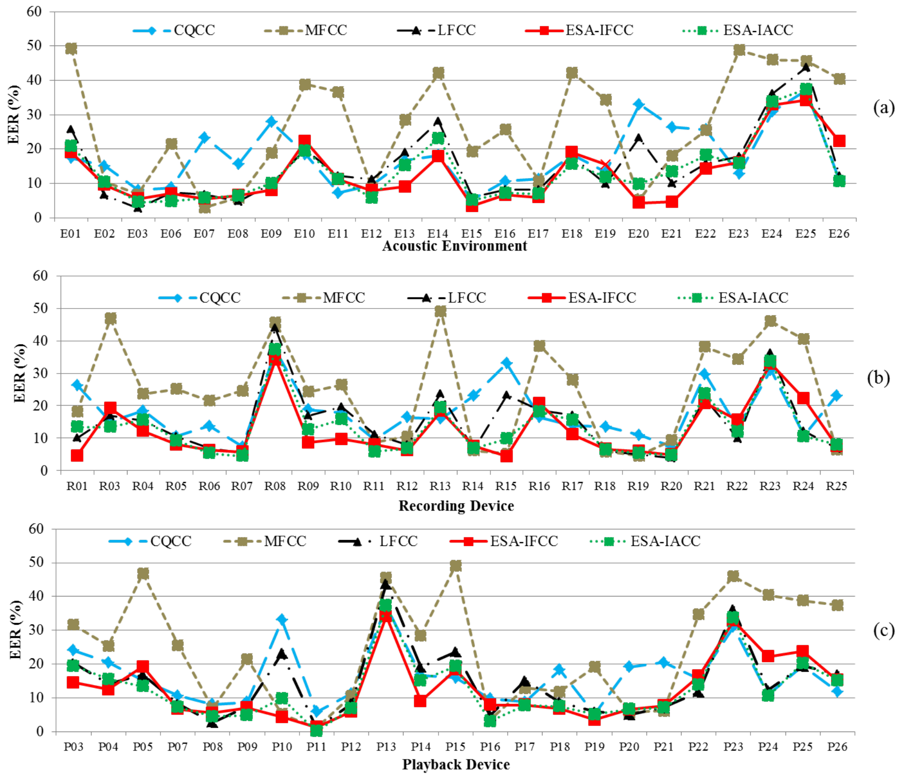


Figure 4.15: Individual % EER for Different Acoustic Environments With CQCC, MFCC, LFCC, ESA-IFCC, and ESA-IACC Feature Sets. After [12].

are 25 and 26 different recording and playback devices denoted by R01-R25 and P01-P26 [6]. Figure 4.15(b) and Figure 4.15(c) shows the detailed % EER for different recording and playback devices with all the feature sets on evaluation dataset. The high-level threats are difficult to detect due to the use of professional audio equipment, such as active studio monitors, studio headphones, etc. to produce replay samples [6]. The proposed feature set perform better in such high-level threat is shown by the highlighted ovals in Figure 4.15. The proposed feature set shows lower % EER for all replay configurations compared to the other feature sets.

The performance evaluation is also shown by the DET curves for CQCC, MFCC, LFCC, and proposed feature sets along with their best fusion results in Figure 4.16 on version 2.0 database. It can be observed that the miss probability of CQCC, MFCC, and LFCC was very high for given FAR which is not a good case for ASV system. There is decrease in miss probability for proposed feature sets on development set as shown in Figure 4.16 (left side), which further reduces when fused with CQCC feature set. For evaluation set, we observe very high FRR for all the feature sets along with proposed feature sets as shown in Figure 4.16 (right side).

However, the proposed feature sets and their score-level fusion with CQCC has low FAR compared to the other feature sets.

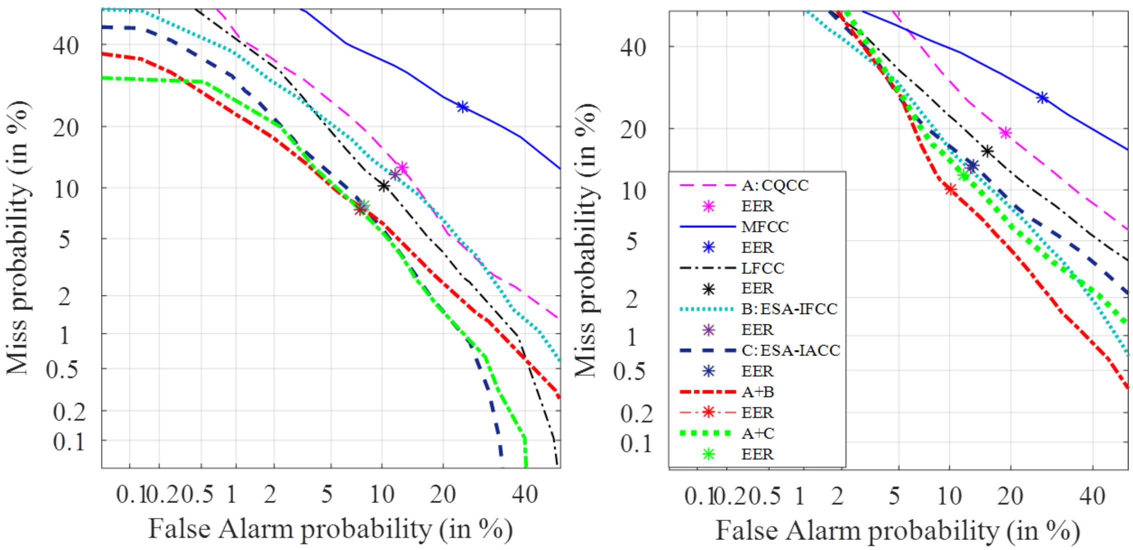


Figure 4.16: The DET Curves for CQCC, MFCC, LFCC, and Proposed Feature Sets Along With Best Score-Level Fusion With Factor α=0.7 (As Per Eq. 6.7) on Development Set (Left Side), and Evaluation Set (Right Side). After [12].

Finally, we have compared our proposed feature sets with the other feature sets that were proposed on the ASVspoof 2017 challenge version 2.0 database. A few studies are reported on the modified database as listed in Table 4.5.

4.4.2 Results on ASVspoof 2015 Challenge

The subband filter must be as wide as possible to include the desired formant modulations. However, narrow enough to exclude the interference of neighboring formants. The center frequencies of the bandpass filters are linearly-spaced and used to extract the component AM-FM signals of the speech segment and then determine the modulations around these center frequencies. Authors have chosen linearly-spaced filterbank for Butterworth filter as opposed to the other frequency scale, such as Mel scale, Equivalent Rectangular Bandwidth (ERB) scale (this is in line with the recent finding reported in [167]).

4.4.2.1 Results on Development Set

Results for MFCC and proposed ESA-IFCC feature set are shown in Table 4.6. From the results, proposed feature set captures speaker-specific information embedded in natural speech (as SS and VC speech does not exactly match the human speech) and hence, there exists differences between natural *vs.* spoofed classes.

Table 4.5: Comparison of Results (in % EER) on ASVspoof 2017 Version 2.0 Challenge Database. After [12]

Feature Set	Classifier	Dev	Eval
CQCC [6] (Post-eval BL)	GMM	9.06	13.74
A: CQCC	GMM	12.81	19.04
LFCC	GMM	10.58	16.62
MFCC	GMM	24.19	26.90
PNCC [157]	GMM	20.78	23.74
QLNCC [157]	GMM	21.81	24.67
CILPR [199]	GMM	19.68	20.66
PSRMS [199]	GMM	33.38	28.16
eCQCC-DA [200]	DNN	13.97	13.38
CQCC [201]	GMM	8.93	12.20
IFCC [201]	GMM	16.20	15.90
DCTILPR [201]	GMM	22.69	14.03
RMFCC [201]	GMM	23.58	20.49
TECC [11]	GMM	9.55	11.73
CF [141]	GMM	-	10.84
CM [141]	GMM	-	10.93
PPWS [147]	GMM	-	10.70
PPWS_max [147]	GMM	-	11.57
PPRFWS_KL [147]	GMM	-	9.97
PPRFWS_LR [147]	GMM	-	9.28
B: Proposed ESA-IACC	GMM	7.99	13.45
C: Proposed ESA-IFCC	GMM	11.84	12.93
A+B (α =0.5(dev) & eval))	GMM	**7.03**	12.11
A+C (α =0.7(dev & eval))	GMM	7.74	**10.12**

- indicates information not found;
BL: Baseline

Table 4.6 indicate that the ESA-IFCC features produce much lower % Equal Error Rate (EER) than the MFCC alone, ESA-IFCC features that are capable to classify genuine *vs.* bonafide speech (i.e., for SS and VC, the features are comparatively different than that for human (genuine/natural) speech).

It was found that linearly-scaled equi-spaced subband filters are more suitable for IF estimation task than the ERB-scaled varying bandwidth subband filters. In the case of gammatone filterbank, the bandwidth increases at higher frequencies, making the estimation of IF less reliable. However, authors have used Butterworth filter that has nonlinear phase that can be approximated as linear over smaller frequency regions. For given 16 kHz sampling frequency, we have avail-

Table 4.6: Results in Terms of EER (%) on Development Dataset Score-Level Fusion as Per Eq. 3.28. After [13]

Feature Set 1	# EER (%) for varying α													Feature Set 2
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	Frequency Range	
	static	6.98	6.72	6.36	6.00	5.69	5.40	**5.24**	5.30	5.54	6.63	8.16		
MFCC	Δ	6.75	6.25	5.84	5.36	4.83	4.42	4.13	3.88	**3.85**	4.27	5.29	100-3000 Hz	ESA-IFCC
	ΔΔ	6.14	5.71	5.29	4.81	4.31	3.98	3.74	**3.70**	3.85	4.56	5.79	13-D	
	static	6.98	6.28	5.67	5.05	4.48	4.17	3.91	**3.80**	4.06	4.78	6.38		
MFCC	Δ	6.75	4.05	2.76	2.27	**2.18**	2.34	2.56	3.28	4.17	5.67	7.47	100-7800 Hz	ESA-IFCC
	ΔΔ	6.14	2.98	2.17	**1.98**	2.12	2.42	3.00	3.63	4.62	5.75	7.18	40-D	
	static	6.98	6.57	6.13	5.67	5.18	4.66	4.16	3.75	**3.45**	3.62	5.43		
MFCC	Δ	6.75	5.58	4.50	3.72	3.14	2.71	2.30	**2.01**	2.02	2.71	6.22	100-7800 Hz	ESA-IFCC
	ΔΔ	6.14	5.10	4.08	3.31	2.82	2.41	2.02	**1.89**	2.00	2.76	6.59	13-D	

Table 4.7: Results in % EER on Evaluation Dataset for Each Spoofing Attack. Both Known and Unknown Attacks, +:Score-Level Fusion. After [13]

Feature Sets	Known Attacks						Unknown Attacks						All Avg.
	S1	S2	S3	S4	S5	Avg.	S6	S7	S8	S9	S10	Avg.	
A:MFCC	2.34	9.57	**0.00**	**0.00**	9.01	4.18	7.73	4.42	0.3	5.17	52.99	14.12	9.15
B:ESA-IFCC	2.68	4.87	**0.00**	**0.00**	12.87	4.08	10.9	2.4	3.57	3.33	**37.37**	**9.514**	**6.79**
A+B	**0.78**	**3.39**	**0.00**	**0.00**	**5.45**	**1.92**	**4.19**	**1.22**	**0.11**	**1.80**	54.73	**12.41**	7.16

able bandwidth of 7800 Hz that is divided into 40 equi-spaced frequency regions of width $(f_H - f_L)/40$ Hz. The phase response around each $(f_H - f_L)/40$ Hz width is mostly found to be linear (as observed in authors recent study reported in [202]).

Furthermore, the score-level fusion of these features was done as per Eq. (3.28) and is shown in Table 4.6. It was observed that for equal weighted fusion of MFCC and ESA-IFCC score, the % EER of MFCC (6.98 %), and ESA-IFCC (5.43 %) reduces to 3.45 % for static features similar pattern was observed for Δ and ΔΔ for higher frequency range of 100-7800 Hz. The contribution of a particular system is decided by the weight of fusion (i.e., α). From Table 4.6, it is observed that most of the system contribution was done with ESA-IFCC feature set as the parameter α most of the times was biased towards ESA-IFCC. Therefore, it can be said that the ESA-IFCC feature set has more contribution in decreasing the % ERR and hence, could classify in a more effective way for SSD. From Table 4.6, the proposed feature set capture the *complementary* information that was not observed from MFCC alone. The % EER, is very less for score-level fusion of the MFCC and ESA-IFCC.

4.4.2.2 Results on Evaluation Dataset

Table 4.7 shows the results for evaluation dataset with known and unknown spoofing attacks. It was observed that SS attacks (S3, S4) were easily detected for known attacks, while S10 (MARY TTS) in unknown attacks was the most difficult task to detect. These results show that performance degrades significantly with unknown attacks. The overall average error rate for known and unknown attacks was 6.79 % for ESA-IFCC and was significantly better than the MFCC (9.15 %) feature set. The score-level fusion (when performed with the fusion factor, $\alpha = 0.8$) gave the overall average EER of 7.16 % due to dominance of S10 unknown attack. However, with score-level fusion of MFCC and ESA-IFCC feature set, other attacks from S1 to S9 (known and unknown attacks) were detected reasonably well.

4.5 Chapter Summary

In this chapter, we discussed two proposed feature sets, namely, ESA-IFCC and ESA-IACC for the SSD task. We presented the results using several experimental evaluation factors, such as the shape of subband filters, frequency scales (ERB, Mel, and linear), the number of subband filters, type of filterbank (Gabor *vs.* Butterworth), etc. The AM-FM features were extracted using ESA-based speech demodulation approach. The ESA approach uses a nonlinear differential operator. This study has shown that ESA method exhibit better demodulation performance for the replay SSD task. The Gabor filterbank is used to obtain subband filtered AM-FM speech signals. We observed that the AM-FM feature set extracted using linearly-spaced Gabor filters gave better performance than that for ERB and Mel scales. The frequency resolution is explicitly related to the number of subband filters in the filterbank. The AM-FM features along with feature normalization performed better compared to the baseline system, and many other features in the literature. The score-level fusion of IA and IF features improve the performance, however, further investigation is required in this direction.

The limitation of this study is to understand the rapid IF fluctuation changes around the center frequency for voice and unvoiced region. In the next Chapter, we extend Chapter 4 work with the generalized TEO by varying the past and future samples instants with a constant integer known as *lag parameter*.

CHAPTER 5

Variable Length Energy Separation Algorithm (VESA)

5.1 Introduction

In the Chapter 4, we discussed our proposed ESA-based Instantaneous Amplitude and Frequency feature sets. In this chapter, we extend our earlier work with the generalized TEO, i.e., by varying the samples of past and future instants with a constant arbitrary integer i also known as *lag parameter or dependency index*, and named it as Variable length Teager Energy Operator (VTEO). The key concept behind VTEO, and its Variable length Energy Separation Algorithm (VESA) is discussed in Section 5.2 and Section 5.3. We compared the VESA-based features with earlier proposed method, i.e., ESA along with Hilbert transform method for SSD task. In particular, we performed experiments on ASVspoof 2017 challenge and BTAS database in Section 5.4. In addition, the experimental results using VTECC feature set ASVspoof 2017 challenge V2.0 database is reported in Section 5.5. Finally, Section 5.6 summarizes the Chapter.

5.2 Basics of VTEO

Variable length Teager Energy Operator (VTEO) is the modified version of the traditional TEO method [165]. TEO involves nonlinear operations on the signal, i.e, square of current sample and multiplication of previous and next sample, i.e., $x(n-1)$ and $x(n+1)$, respectively. The key motivation for VTEO is the speech signal carries dependencies (local *vs.* distant) in the sequence of samples of speech signal. Thus, instead of considering only immediate past $x[n-1]$ and immediate future $x[n+1]$, VTEO considering i^{th} past and i^{th} future samples. In VTEO algorithm, the number of samples incorporated in energy estimation can be varied up to i past, and i future samples, i.e., $x(n-i)$ and $x(n+i)$, instead of only two adjacent samples as in TEO [203]. VTEO gives flexibility to select these samples to estimate the running estimate of energy required to generate the signal [204]. VTEO gives us a good measure of the energy of the oscillating signal, when the sampling rate of the signal is greater than $8i$ times the frequency of oscillation in the signal [203]. VTEO brings out *hidden* dependencies and dynamics of the signal [203]. For discrete-time signal, $x[n] = Acos(\omega n + \phi)$, the samples of the same signal shifted in time by index i, w.r.t present sample, can be expressed with an assumption for i>n, $x(n-i) = 0$ as:

$$x(n+i) = Acos(\omega(n+i) + \phi), \tag{5.1}$$

$$x(n-i) = Acos(\omega(n-i) + \phi). \tag{5.2}$$

When we multiply above equations, we obtain,

$$x(n+i)x(n-i) = A^2cos(\omega(n+i) + \phi)cos(\omega(n-i) + \phi), \tag{5.3}$$

$$x(n+i)x(n-i) = [Acos(\omega n + \phi)]^2 - A^2sin^2 i\omega. \tag{5.4}$$

On high sampling rates, it result to VTEO and is given as Eq. (5.5):

$$E_n = \{\Psi_{DI}\{x(n)\}\} = x^2(n) - x(n-i)x(n+i) \approx i^2A^2\omega^2, \tag{5.5}$$

where $i^2A^2\omega^2$ is instantaneous estimate of signal's energy multiplied by i^2, and referred to as VTEO for the dependency index (DI), i, which is expected to give running estimate of signal's energy [204,205]. To estimate the individual contribution of amplitude, $a[n]$, and frequency, $\omega[n]$, of signal, Maragos *et.al* developed an *Energy Separation Algorithm (ESA)* that uses nonlinear energy operator (i.e., in TEO framework) to track the instantaneous energy of the source generating the AM-FM signal, and separate it into its amplitude and frequency components [51, 57]. The ESA was developed to demodulate a speech signal into Amplitude Envelope (AE) and Instantaneous Frequency (IF). According to Kaiser, energy in a speech signal is a function of both amplitude and frequency [206]. However, ESA is applied to single speech resonance, while the speech signal itself is a multi-component signal, being the sum of several resonances. Hence, there is a need to isolate resonances by suitable bandpass filtering. The advantage of VTEO over TEO lies in the superior localization, and tracking instantaneous fluctuations (if any) of the energy at a given instant of time, and it also brings out the *hidden* dependencies and dynamics of the signal w.r.t. distantly located speech samples than only the immediate adjacent samples. The VTEO has a good measure of

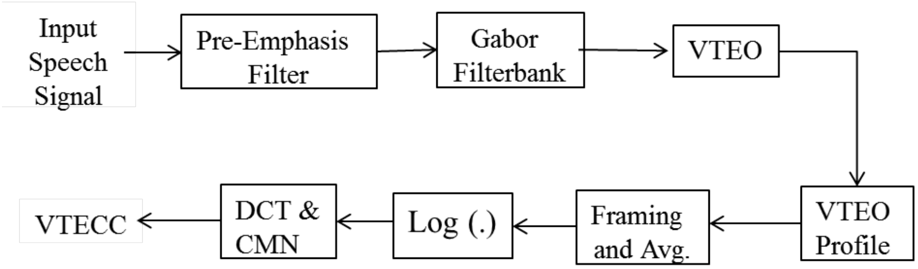


Figure 5.1: Block Diagram of Proposed Variable Length Teager Energy Cepstral Coefficients (VTECC). After [16].

the energy, when the sampling rate is greater than 8i times the frequency of the oscillation of signal [203].

5.2.1 Feature Extraction Process

The block diagram of Variable length Teager Energy Cepstral Coefficients (VTECC) feature set is shown in Figure 5.1. VTECC is an extension of our recent study reported in [10, 11]. VTECC is found to perform better for SSD task, synthetic, and voice converted speech (SS and VC) signal as per our recent work done on the ASVspoof 2015 challenge database [10]. The VTECC was computed by first filtering the speech signal through a dense non-constant-Q Gammatone filterbank for robust speech recognition in [207, 208]. The input speech signal is given to the filterbank to obtain N number of subband signals [51, 52]. We have used linearly-spaced Gabor filterbank to have almost equal bandwidth to cover the entire frequency range [14, 27, 29]. Furthermore, these subband filtered signals are given as input to the TEO block to estimate the energy profile of each subband filtered signals. These TEO profiles are passed through the frame-blocking, and averaging using a short window length of 20 ms with a shift of 10 ms followed by logarithm operation to compress the data. The Discrete Cosine Transform (DCT) is then applied for energy compaction, and retained first few DCT coefficients in order to obtain VTECC feature set, followed by their Δ and $\Delta\Delta$ feature vector to obtain higher-dimensional static plus dynamic feature vector. From the earlier studies on replay SSD task, we found that the higher frequency regions are more useful along with Cepstral Mean Normalization (CMN) technique. Hence, VTECC feature set is extracted using pre-emphasis filter and CMN technique [14, 27].

We observed the Teager energy traces of the speech segment considered for natural (blue line) and replay (red line) as shown in Figure 5.2. We can see that for the segment of replay speech very high (impulse-like) energy traces are obtained when compared to the segment of natural speech. In addition, we also observed

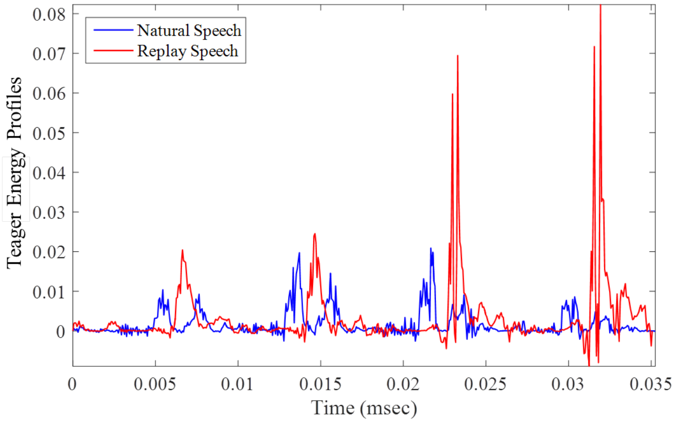


Figure 5.2: Teager Energy Traces of the Natural (Blue Line) and Replay (Red Line) Speech Segment. After [16].

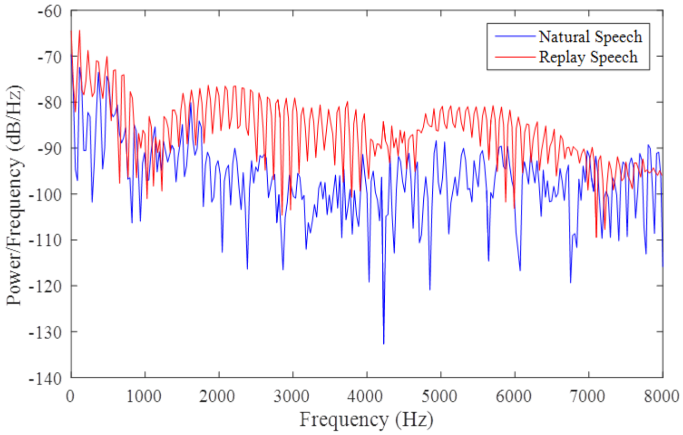


Figure 5.3: Power Spectral Density (PSD) of Teager Energy Traces of the Natural and Replay Speech Segment. After [16].

the Power Spectral Density (PSD) for Teager energy traces of natural and replay speech segment as shown in Figure 5.3. The variation at each frequency component for Teager energy traces of replay segment (red line) are more smooth compared to that of Teager energy traces of natural segment (blue line).

5.2.2 Analysis of Variable length Teager Energy Profiles

The VTEO profiles corresponding to DI= 1 to 10 are shown in Figure 5.4. The blue line corresponds to natural Teager energy profiles, and red line to replay speech signals. For the initial DIs, i.e., from DI= 1 to 2 for replay signal, we cannot see the profiles clearly as they are all merged around the glottal closure instants (GCI's). After DI=2 the replay signal profiles start to show the Teager energy profiles similar to the natural signal. Later after DI=6, more fluctuations and bumps are observed in replay signal whereas it is reduced for the case of natural signal as

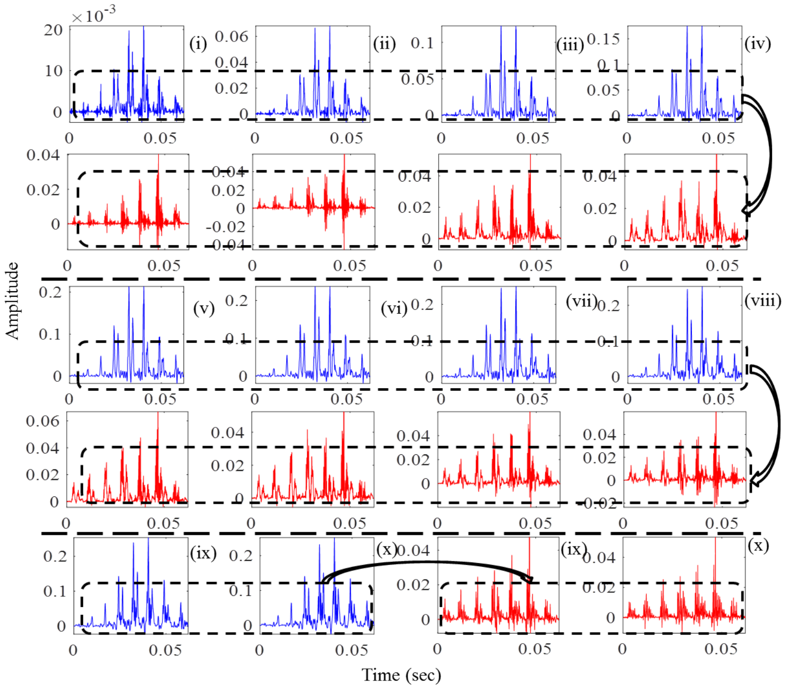


Figure 5.4: Teager Energy Profiles with Varying the DI from DI=1 to 10. Blue and Red Teager Energy Profiles Corresponds to Natural and Replay Signal. Highlighted Regions Show the Difference in Teager Energy Profiles. After [16].

we increase the DI after 6. According to the results shown in experimental result section, with DI=5 the replay signals are detected and classified well compared to other DIs.

5.2.3 Spectral Energies of Variable length Teager Energy

Figure 5.5 show the spectral energy corresponding to each DI obtained from Variable length Teager energy. The spectral energies here is shown for the natural speech signal. It can be observed from the Figure 5.5 that with every DI we find some differences corresponding to the first DI (shown by highlighted circles). With DI=5, we observe more spectral energy differences in lower as well as in higher frequency regions. This spectral energy changes corresponding to other DI helps to detect and classify it from the natural signal. This can also be observed from the results obtained from all the DIs reported in Section 5.5, where we obtained relatively lower % EER for DI=5.

5.3 Basics of VESA

In this Section, we propose to exploit VTEO to track the modulation energy and estimate the instantaneous amplitude and frequency of AM-FM signal, and refer

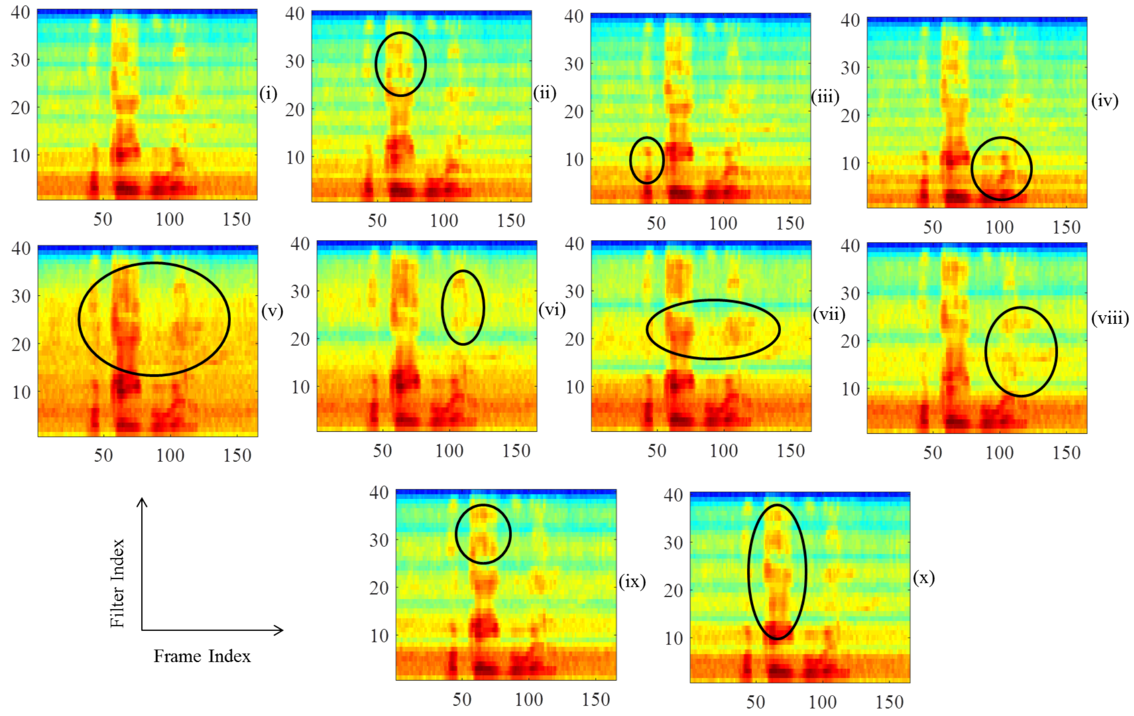


Figure 5.5: Spectral Energy Densities Obtained from Variable Length Teager Energy with Varying the DI from (i) (DI=1) to (x) (DI=10). Highlighted Circles Show the Differences in Spectral Energies. After [16].

it to as VESA. The IF $\omega[n]$ and AE $a[n]$ at any time the instant of the AM-FM modulated signal $x[n]$ is given by [14,28]:

$$IA_{VESA} = a_i[n] \approx \frac{2\Psi_{DI}\{x[n]\}}{\sqrt{\Psi_{DI}\{x[n+1]-x[n-1]\}}}, \tag{5.6}$$

$$IF_{VESA} = \omega_i[n] \approx arcsin\sqrt{\frac{\Psi_{DI}\{x[n+1]-x[n-1]\}}{4\Psi_{DI}\{x[n]\}}}. \tag{5.7}$$

Eq. (5.6) and Eq. (5.7) reduces to original ESA algorithm, when DI=1, for $\Psi_{DI}\{x(n)\} = TEO(x(n))$. The frequency estimation part assumes that $0 < \omega_i[n] < \frac{\pi}{2}$ because the computer implementation of $arcsin(u)$ function assumes that $|u| < \frac{\pi}{2}$. Thus, discrete ESA can be used to estimate IF $< 1/4$ of sampling frequency of signal [57]. The IF is modeled as the superposition of slow and fast-varying components. The slow-varying component models the average of formant frequency, and the fast-varying component models frequency variations around the formant frequency. This instantaneous energy can be decomposed using Variable length Energy Separation Algorithm (VESA), and estimate the Instantaneous Amplitude and Instantaneous Frequency (IA-IF) of a speech signal.

5.3.1 Proposed VESA-IFCC Feature Set

Figure 5.6 shows the block diagram of proposed VESA-IFCC feature set. Here, the input speech signal is first split into N frequency subband signals. The ESA is applied using VTEO with various dependency index (DI) ($DI = 1$ to 10) onto each N bandpass (subband) filtered signals to estimate corresponding IAs and IFs. Furthermore, we have discarded the IA and estimated only IF for each of the narrowband components in order to emphasize the spectral envelope of genuine *vs.* replayed speech. The IFs are segmented into overlapping short (segmental) frames of 20 ms duration, shifted by 10 ms, and the temporal average is estimate to obtain N-dimensional IFCs for every frame. The redundancy among IFCs is exploited to obtain a low-dimensional representation by employing DCT that has energy compaction property and thus, retaining first few DCT coefficients that are referred to as Instantaneous Frequency Cosine Coefficients (IFCC). The IFCC along with their delta and double-delta features were also appended resulting in higher-dimensional feature set denoted as VESA-IFCC. Algorithm 1 shows the procedure for extracting VESA-IFCC features. The Amplitude and Frequency Modulation (AM-FM) features estimate using different demodulation techniques, namely, HT and ESA are discussed in the next Section.

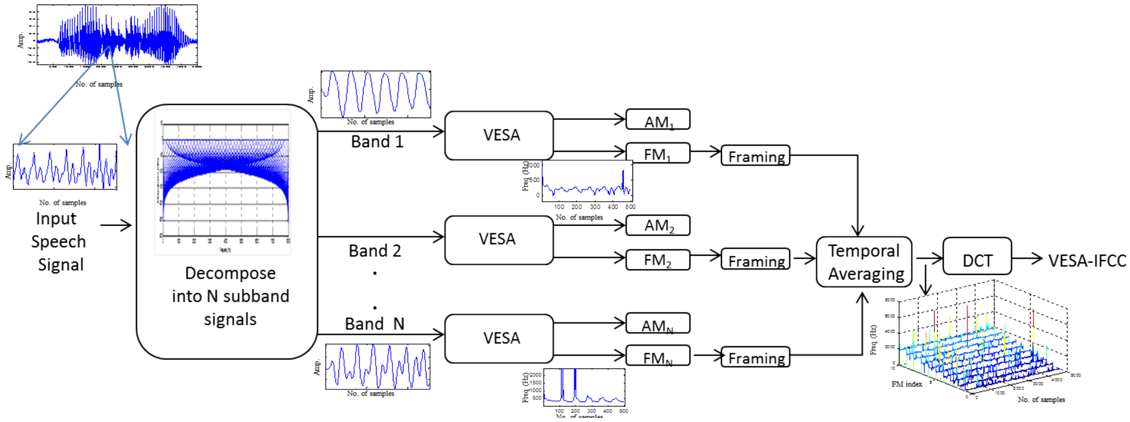


Figure 5.6: Schematic Diagram to Estimate Proposed VESA-IFCC Feature Set. The 3-D Plot Before DCT Corresponds to 5000 Samples. After [28].

5.3.2 Hilbert Transform (HT)

The Hilbert transform estimate amplitude envelope, and frequency function of a monocomponent signal with certain conditions, namely, the frequency variation should not be large (i.e., the signal should be narrowband), and the amplitude variations should not be large [183, 209]. Let $x_a(t)$ be the analytic signal corre-

Algorithm 1 The VESA-IFCC Feature Extraction from the Speech Signal. After [28]

```
$x(n)$= Speech signal.
Consider an $N$-channel filterbank having linearly-spaced Butterworth filters in time-domain.
for $i$=1 to $N$ do
    Perform narrowband filtering of $x(n)$ through $i^{th}$ filter; $x_i(n)$.
    Estimate VTEO profile from $x_i(n)$ as in Eq. (5.5)
    Estimate VESA and extract IF $\omega_i(n)$ as in Eq. (5.7).
end for
Segment $\omega_i(n)$, $i = 1, 2, ....., N$ into short-time frames of duration as 20 ms, shifted by 10 ms.
Average IF for each frame to obtain $N$-dimensional IFCs.
Apply DCT on VESA-IFCs and retain first few coefficients to get VESA-IFCCs.
Append VESA-IFCCs with their first and second-order derivatives.
```

sponding to the real signal, $x(t)$, then $x_a(t)$ is given by:

$$x_a(t) = x(t) + jHx(t) = x(t) + j\hat{x}(t); \tag{5.8}$$

where quadrature signal $\hat{x}(t)$ is the Hilbert transform $Hx(t)$ of $x(t)$. The $\hat{x}(t)$ can be equivalently defined through the Fourier transform as:

$$\hat{X}(\omega) \longleftrightarrow -jsgn(\omega)X(\omega) = \begin{cases} -jX(\omega), & \omega > 0, \\ +jX(\omega), & \omega < 0. \end{cases} \tag{5.9}$$

and

$$X_a(\omega) = \begin{cases} 2X(\omega), & \omega > 0, \\ 0, & \omega < 0. \end{cases} \tag{5.10}$$

with $\hat{X}(\omega)$, and $X(\omega)$ being the Fourier transform of $\hat{x}(t)$ and $x(t)$, respectively. For a given signal, $x(t) = a(t)cos(\phi(t))$, where $cos(\phi(t))$ denotes the Hilbert fine structure, $\phi(t)$ is instantaneous phase, and is defined as $\phi(t) = arctan\left(\frac{\hat{x}(t)}{x(t)}\right)$. The IA, $a_h(t)$ and IF, $\phi_h'(t)$ are derived from the analytic signal as:

$$a_h(t) = \sqrt{x^2(t) + \hat{x}^2(t)}, \tag{5.11}$$

$$\phi_h'(t) = \frac{d}{dt}(\phi(t)). \tag{5.12}$$

As discussed in earlier Chapter 4, we compared ESA-based technique with VESA as they obey similar mathematical structure with having a difference in *lag* *parameter or Dependency Index (DI).*

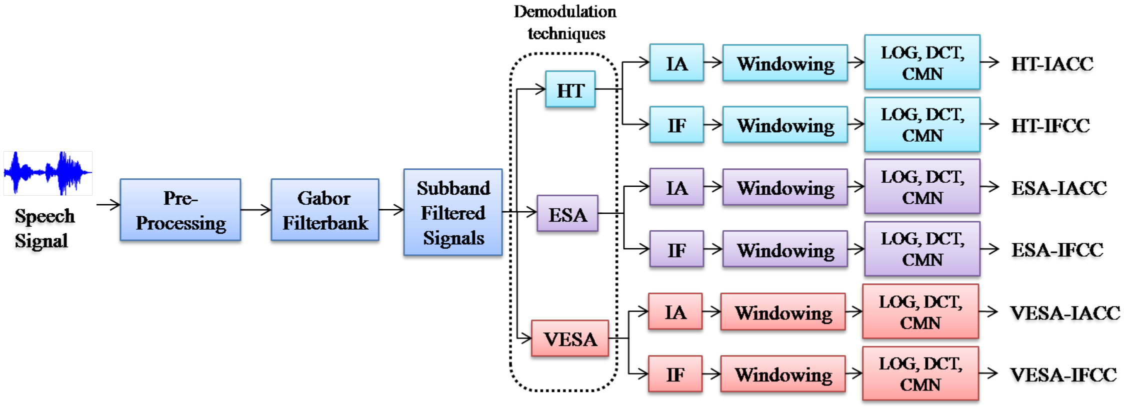


Figure 5.7: Block Diagram for Feature Extraction of IA and IF-Based Features Using HT, ESA, and VESA. After [15].

The block diagram of speech demodulation technique-based features are shown in Figure 5.7. The IA and IF component-based feature sets proposed in the earlier studies are reported in [13, 14, 27, 28]. Initially, signal is passed through the pre-emphasis filter, and then passed through the filterbank to obtain N number of subband signals [51–53]. We used linearly-spaced Gabor filterbank to have almost equal bandwidth to cover the entire time-frequency range [14,27,29]. Furthermore, these subband filtered signals are given as input to the HT, ESA, and VESA block to estimate corresponding IA and IF components. These individual IA and IF components are passed through the frame-blocking and averaging using a short window length of 20 ms with a shift of 10 ms followed by logarithm operation to compress the data. The Discrete Cosine Transform (DCT), and Cepstral Mean Normalization (CMN) technique is then applied for energy compaction, and retained first few DCT coefficients to obtain HT, ESA, and VESA-based IA and IF Cepstral Coefficients, (i.e., IACC and IFCC), followed by their Δ and $\Delta\Delta$ as dynamic features to obtain higher-dimensional feature vector.

The spectral energy density obtained from all the three speech demodulation techniques are shown in Figure 5.8 for a time-domain speech signal (a). The corresponding spectral energy for HT is shown in Figure 5.8(b), for ESA it is shown in Figure 5.8(c), and for VESA, it is shown in Figure 5.8(d). The highlighted dotted box in the Figure 3 shows the spectral differences for all the three different techniques. With VESA-based spectral energy, it can be observed that the high resolution for the harmonics, and frequency bands in the lower frequency region is obtained.

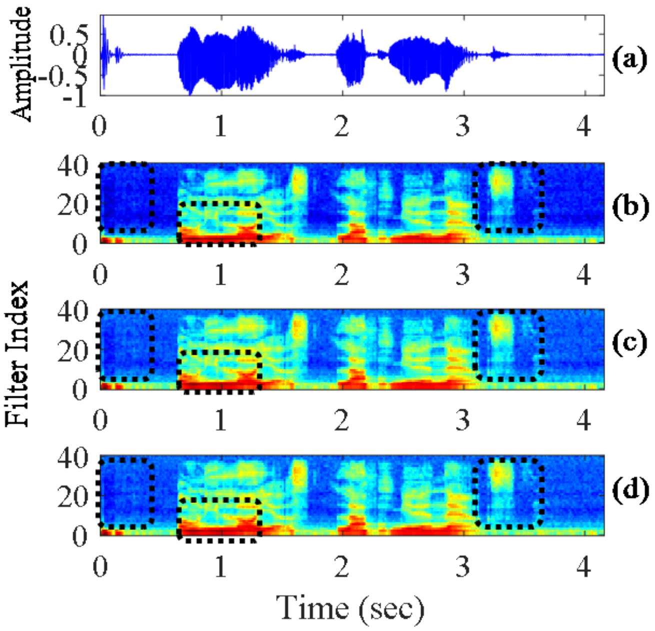


Figure 5.8: (a) Time-Domain Speech Signal, and its Corresponding Spectral Energy Densities Using (b) HT, (c) ESA, and (d) VESA with DI=2. Dotted Box Indicates the Spectral Energy Differences Obtained from Corresponding Demodulation Techniques.

5.4 Experimental Results Using VESA

5.4.1 Results on ASVspoof 2017 V1.0 Challenge

5.4.1.1 Results on Development Set

The results with varying the DI from DI=1 to 4 on development (dev) set of the proposed feature set VESA-IACC are shown in Table 5.1. The results of our recently proposed VESA-IFCC are also shown Table 5.1 [28]. The VESA-IACC feature set obtained an EER of 6.12 % with DI=1, whereas with VESA-IFCC the EER is 6.61 % for DI=2. Since VESA-IACC and VESA-IFCC capture distinct information of amplitude and frequency to explore possible complementary information captured by them, their score-level fusion is done. Thus, to explore this individual information of both the feature sets, we have fused these features for all the four DIs. From Table 5.1, it can be observed that after score-level fusion of VESA-IACC and VESA-IFCC for each DI, the EER is reduced than that for individual feature sets indicating that these two feature sets indeed capture complementary information than the individual feature sets alone. The best lower EER was obtained with the fusion of features at DI=4 resulting in the reduced EER of 0.19 % from 7.18 % (for VESA-IACC), and 6.63 % (for VESA-IFCC) feature sets clearly demonstrating the potential of idea of exploring DI in TEO [203].

Table 5.1: Results (% EER) on Dev Dataset of Proposed Feature Set on Various Dependency Index (DI). After [14]

DI	VESA-IACC	VESA-IFCC [28]	Score-Level Fusion
1	**6.12**	7.65	1.72
2	7.44	**6.61**	0.33
3	7.83	6.65	0.26
4	7.18	6.63	**0.19**

5.4.1.2 Results on Evaluation Set

Similarly, we extracted proposed feature sets on evaluation (eval) set as done on dev set. We varied the DI from 1 to 4 for both VESA-IACC and VESA-IFCC feature sets. The lower EER was obtained on VESA-IACC is with DI=2 of 11.94 %, while with VESA-IFCC feature set, we got EER of 11.79 % with DI=4. The best lower EER was obtained with the score-level fusion of VESA-IACC, and VESA-IFCC feature set reducing the EER to 7.11 % on DI=4 from the individual EER of 12.27 % (for VESA-IACC), and 11.79 % (for VESA-IFCC) feature sets.

Table 5.2: Results on Eval Dataset of Proposed Feature Set on Various Dependency Index (DI) in VTEO. After [14]

DI	VESA-IACC	VESA-IFCC	Score-Level Fusion
1	12.08	16.16	9.11
2	**11.94**	13.46	7.56
3	12.09	12.34	7.15
4	12.27	**11.79**	**7.11**

5.4.1.3 Results of Score-Level Fusion

To explore the possible complementary information present in other feature sets, namely, CQCC, LFCC, and MFCC, we have used the scores of those features and fused them at the score-level with VESA-IACC and VESA-IFCC as shown in Table 5.3.

The score-level fusion is performed for every DI on both dev and eval datasets. The score-level fusion with CQCC indeed helped to reduce the EER for each DI. On dev set, the lower EER of 3.99 % was obtained when fused with DI=1 in VESA-IACC feature set while with DI=4, the fusion of VESA-IFCC and CQCC gave lower EER of 3.28 %. The next fusion was done with LFCC feature set, the score-level fusion of VESA-IACC and LFCC did not reduce the EER, whereas the fusion with VESA-IFCC reduced the EER to 0.38 % with DI=4. The VESA-IACC feature

Table 5.3: Results (in % EER) on Score-Level Fusion of CQCC, LFCC, and MFCC with Various Dependency Index (DI) on Dev and Eval Dataset. After [14].

		Dependency Index (DI)							
		VESA-IACC				**VESA-IFCC**			
	DI	1	2	3	4	1	2	3	4
Dev	CQCC	**3.99**	4.27	4.40	4.39	5.32	3.75	3.64	**3.28**
	LFCC	**6.12**	7.44	7.38	7.18	2.23	0.74	0.56	**0.38**
	MFCC	**4.06**	4.36	4.39	4.31	3.21	**1.42**	2.26	1.58
Eval	CQCC	11.18	**11.13**	11.28	11.49	16.16	13.46	12.34	**11.79**
	LFCC	12.08	**11.94**	12.09	12.27	10.32	8.44	7.84	**7.93**
	MFCC	12.08	**11.94**	12.09	12.27	16.16	13.46	12.34	**11.79**

set was extracted with the linear scale and LFCC also uses the linear scale and hence, possibly the score-level fusion did not reduce the EER as both are magnitude spectrum-based features and thus, may not carry much complementary information. For most of the DIs, the lower EER obtained was the same as that of the VESA-IACC feature set. On the other hand, the fusion of VESA-IFCC and LFCC also uses linear scale, however, they carry the complementary information of magnitude and phase spectra because of which EER is reduced. Finally, we fused our feature sets with MFCC obtaining an EER of 4.06 % with VESA-IACC for (DI=1) and 1.42 % for (DI=2) with VESA-IFCC features.

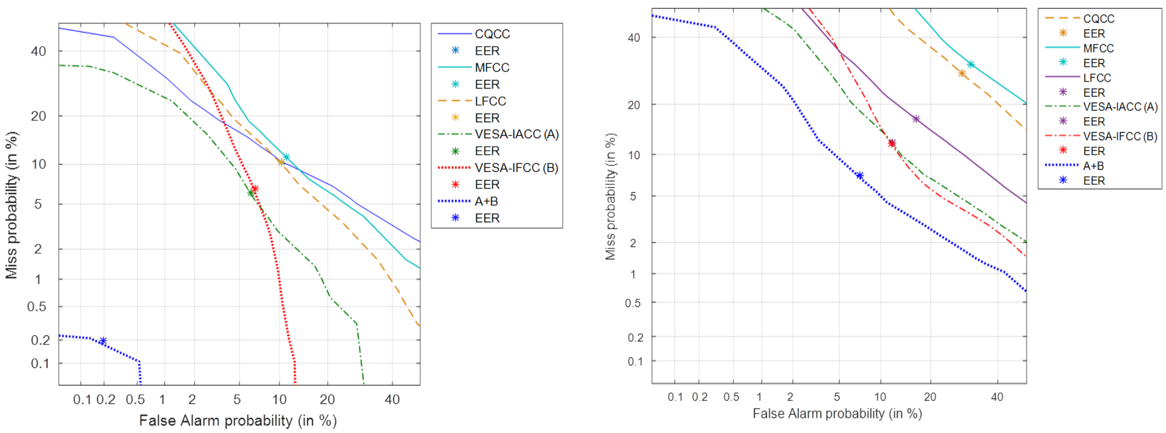
Similarly, the score-level fusion was performed on the eval dataset. There was a reduction in the EER when fused with VESA-IACC (for DI=1) and CQCC resulting in 11.13 %, whereas the fusion of VESA-IACC (for DI=2) with LFCC and MFCC obtained reduced EER of 11.94 % as shown in Table 5.3. On the other hand, the score-level fusion of VESA-IFCC with CQCC and MFCC did not reduce the EER for all the DIs. While the score-level fusion of VESA-IFCC (with DI=4) and LFCC reduce the EER from the individual system to 7.93 %. Table 5.4 shows the final results of our proposed feature set. The organizers of ASV Spoof 2017 Challenge provided CQCC feature set with GMM classifier as the baseline system. In this Section, we considered CQCC, and LFCC as two distinct baselines systems. The proposed feature set was extracted with linearly-spaced Gabor filterbank and thus, to compare results obtained with proposed features set, we consider LFCC as the second baseline. At last, we used one of the well known MFCC feature set to compare our results. The EER of all the feature sets, namely, CQCC, LFCC, and MFCC are high on both dev and eval sets. The EER for CQCC (baseline system) gave an EER of 10.21 % and 28.48 % on dev and eval sets, respectively.

The VESA-IACC and VESA-IFCC feature sets individually performed better than the CQCC, LFCC, and MFCC feature sets. The best results are obtained with the score-level fusion of VESA-IACC and VESA-IFCC resulting in the lower EER of 0.19 % on dev set and 7.11 % on eval set.

Table 5.4: Final Results (in % EER) on Dev and Eval Dataset of ASVspoof 2017. After [14]

Feature Set	Dev	Eval
CQCC (Baseline)	10.21	28.48
LFCC	10.58	16.62
MFCC	11.21	31.30
A:VESA-IACC	6.12	11.94
B:VESA-IFCC	6.61	11.79
A+B	**0.19**	**7.11**

+: score-level fusion



(a)

(b)

Figure 5.9: DET Curves on Dev and Eval Datasets. (a) The Individual DET Curves on Dev Set of CQCC, MFCC, LFCC, VESA-IACC (A), VESA-IFCC (B), and Score-Level Fusion A+B, and (b) Similar DET Curves on Eval Set.

The performance is also shown by the DET curve in Figure 5.9(a) for dev set and Figure 5.9(b) for eval set for CQCC, MFCC, LFCC, VESA-IACC, and VESA-IFCC feature sets along with score-level fusion of VESA-IACC, and VESA-IFCC. On dev and eval sets, score-level fusion of VESA-IACC and VESA-IFCC are clearly distinct at *all* the operating points of the DET curve and have a lower false alarm and miss probabilities on the DET curve compared to the CQCC, LFCC, and MFCC feature sets alone.

5.4.2 Results on ASVspoof 2017 V2.0 Database and BTAS 2016

The results with varying the lag parameter also called as *dependency index* (DI) from DI=1 to 4 on development set for VESA-IACC, and VESA-IFCC feature sets on both the databases (i.e., (a) ASVspoof 2017 challenge v2.0, and (b) BTAS 2016 competition are shown in Figure 5.10). It can be observed that both IA and IF-based feature sets gave lower EER at DI=2 on both the databases. On ASVspoof 2017 challenge v2.0 database, the EER varies from 7.31 % to 8.04 % for IA-based features whereas for IF-based features, it varies from 20.36 % to 25.27 %. On the other hand, for BTAS 2016 database, the variation is from 2.31 % to 2.59 % and from 5.3 % to 6.14 % for IA and IF-based features, respectively. Hence, for further set of experiments reported in this Chapter, VESA-based features were extracted using DI=2.

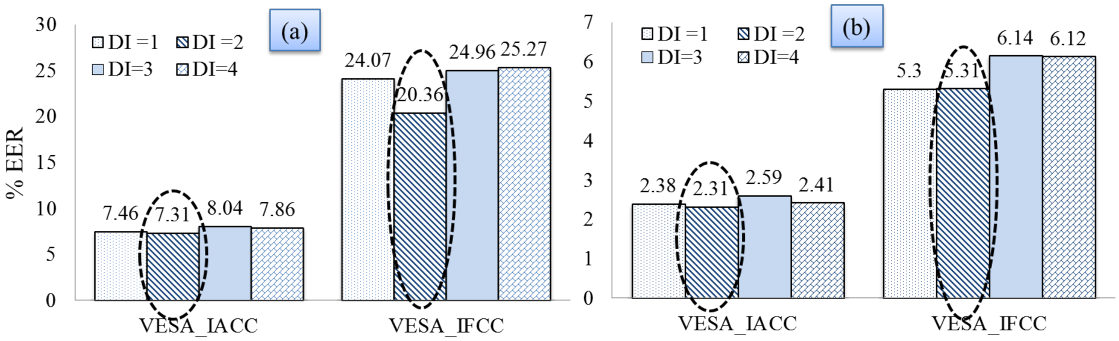


Figure 5.10: Results for Varying DI from DI=1 to 4 on Development Set of (a) ASVspoof 2017 Challenge V2.0, and (b) BTAS 2016 Competition Database. After [15].

5.4.3 Results on Development and Evaluation Sets

Results on all the speech demodulation techniques for both ASVspoof 2017 challenge v2.0 and BTAS 2016 database are reported in Table 5.5 and Table 5.6, respectively. It can be observed that on development set, HT-based features gave lower EER, whereas for evaluation set, VESA-based features gave better performance than the other two demodulation techniques. However, on BTAS database, the results varies for all the speech demodulation techniques with very less differences in % EER. The advantage of VESA over ESA lies in its localization and approximation to track the instantaneous fluctuations (if any) of the energy at a given instant of time. The VESA brings out the *hidden* dependencies, and dynamics of the signal w.r.t. distantly located speech samples than *only* immediate adjacent samples.

Table 5.5: Results (in % EER) of IACC and IFCC Feature Sets Using HT, ESA, and VESA on ASVspoof 2017 Challenge V2.0 Database. After [15]

	IACC		**IFCC**	
	Dev	**Eval**	**Dev**	**Eval**
HT	**7.16**	12.58	**18.86**	30.18
ESA	7.99	13.45	24.07	19.87
VESA (DI=2)	7.31	**12.57**	20.36	**19.10**

Table 5.6: Results (in % EER) of IACC and IFCC Feature Sets Using HT, ESA, and VESA on BTAS 2016 Database. After [15]

	IACC		**IFCC**	
	Dev	**Eval**	**Dev**	**Eval**
HT	**2.26**	**3.96**	5.26	**7.46**
ESA	2.36	4.31	**5.08**	9.23
VESA (DI=2)	2.31	4.73	5.31	9.13

5.4.4 Results of Score-Level Fusion

To obtain the possible complementary information between two feature sets, we used score-level fusion of two feature sets obtained from the same demodulation techniques. For example, the IA and IF components estimated from HT-based method are fused together in order to obtain the reduced EER and gave good performance. It can be observed from Table 5.7 that with score-level fusion, on both the databases, we reduced the EER from its individual EER. We compared our ASVspoof 2017 challenge v2.0 results with the baseline system of the same database, i.e., CQCC feature set. The baseline system gave EER of 12.81 % and 19.04 % on development and evaluation set, respectively. The best EER obtained on development set is with HT-based method giving an EER of 5.91 %, and on evaluation set, the lower EER is obtained with VESA-based technique resulting in 11.45 %.

Table 5.7: Results (in % EER) of Score-Level Fusion of IACC and IFCC Feature Sets Using HT, ESA, and VESA on ASVspoof 2017 V2.0 and BTAS 2016 Database. After [15]

	ASVspoof 2017 v2.0		**BTAS 2016**	
	Dev	**Eval**	**Dev**	**Eval**
CQCC (Baseline)	12.81	19.04	–	–
HT	**5.91**	12.13	**2.26**	**3.93**
ESA	7.72	12.17	2.36	4.31
VESA (DI=2)	6.99	**11.45**	5.31	4.73

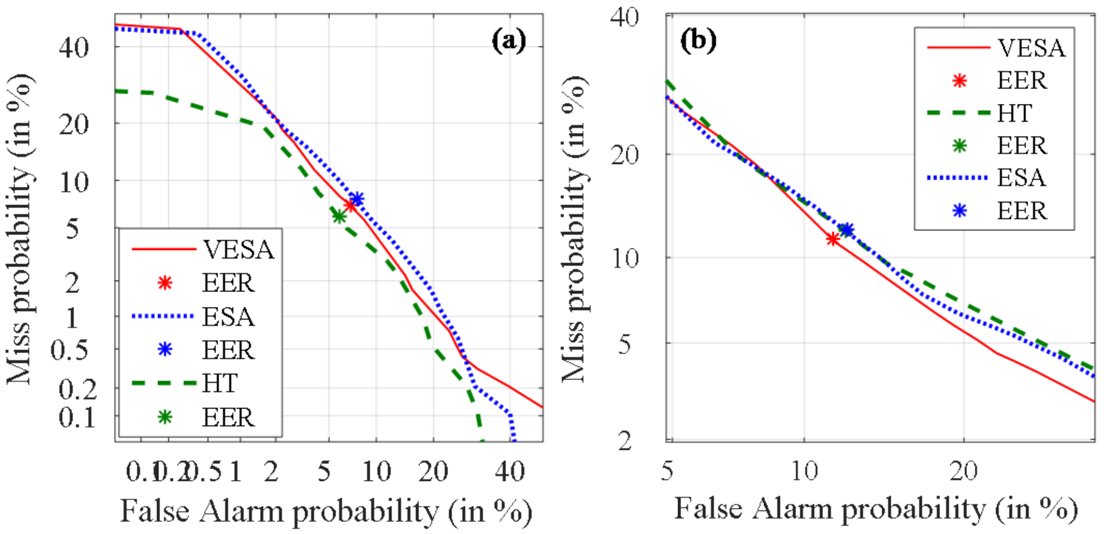


Figure 5.11: DET Curves of Score-Level Fusion of IACC and IFCC on (a) Dev, and (b) Eval Set of ASVspoof 2017 Challenge V2.0 Database. After [15].

Table 5.8 shows the performance on evaluation set in % HTER on BTAS 2016 database (with a single system) and compared our results with the baseline system. The baseline system gave an % HTER of 6.87 % and the best performance of our speech demodulation technique obtained an % HTER of 3.17 % with IA component obtained from HT-based technique. The performance is also shown by the (DET) curve in Figure 5.11 on (a) development, and (b) evaluation set of ASVspoof 2017 challenge v2.0 database. The DET curves are shown only for the score-level fusion of IA and IF components of individual demodulation techniques, i.e., HT, ESA, and VESA, respectively. It can be observed from the DET curves for development set that the HT-based technique gave lower EER with less miss probability and false alarm rate. However, for the evaluation set, the HT technique did not perform well and with VESA method, it performed better. This fluctuations in the performance brings out more generalized countermeasures for SSD task. Note: We have not shown the DET curves for BTAS 2016 database results because the

Table 5.8: % HTER for Eval Set of BTAS 2016

System Used	% HTER
Baseline	6.87
HT-IACC	**3.17**
HT-IFCC	6.74
ESA-IACC	3.64
ESA-IFCC	7.59
VESA-IACC (DI=2)	4.06
VESA-IFCC (DI=2)	7.14

score-level fusion did not reduce the EER from the individual IA-based results.

5.5 Experimental Results Using VTECC

The experiments are performed on ASVspoof 2017 V2.0 challenge database.

5.5.1 Results with Varying Dependency Index (DI)

The results with varying the DI from DI=1 to DI=10 on development set of the proposed VTECC feature set is shown in Figure 5.12. The VTECC feature set obtained an EER of 9.55 % with DI=1, whereas with DI=5 the EER is reduced to 6.52 % which is relatively least % EER among all the DIs. This is because of the spectral energies obtained with DI=5 has more energy as observed and discussed in Figure 5.5. Hence, for further set of experiments, we considered DI=5 for VTEO computation.

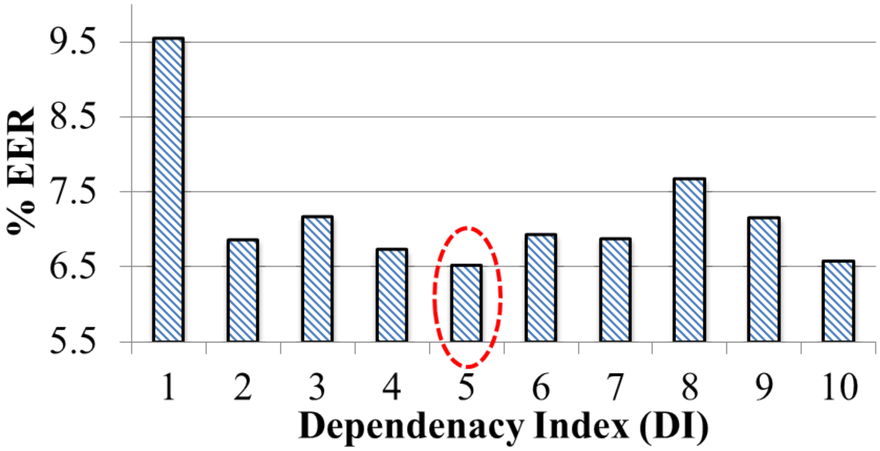


Figure 5.12: Bar Graph Plot with Varying the DI on Development Set. Highlighted Circle Indicates the Least % EER. After [16].

We compared our results with state-of-the-art features, such as CQCC, LFCC, and MFCC. The results obtained from these feature sets for both development and evaluation sets are reported in Table 5.9. Here, the CQCC feature set which is baseline system is extracted in cepstral-domain whereas in actual baseline system, the organizers used log-energy coefficients [6].

The histogram plots of log-likelihood scores obtained from Gaussian mixtures corresponding to (a) CQCC, (b) LFCC, (c) MFCC, and (d) VTECC are shown in Figure 5.13 for development set. It can be observed that with VTECC feature set, the LLR scores of both natural and replay are distributed more resulting in

Table 5.9: Final Results on Dev and Eval Set. After [16]

Feature Set	Dev	Eval
CQCC (Baseline)	12.75	18.97
LFCC	10.31	15.73
MFCC	23.80	26.62
VTECC	6.52	11.93
CQCC+VTECC	**5.85**	**10.94**
LFCC+VTECC	6.52	11.93
MFCC+VTECC	6.52	11.67

Proposed VTECC is computed with DI=5

lower % EER as compared to the distribution obtained from other feature set on development set.

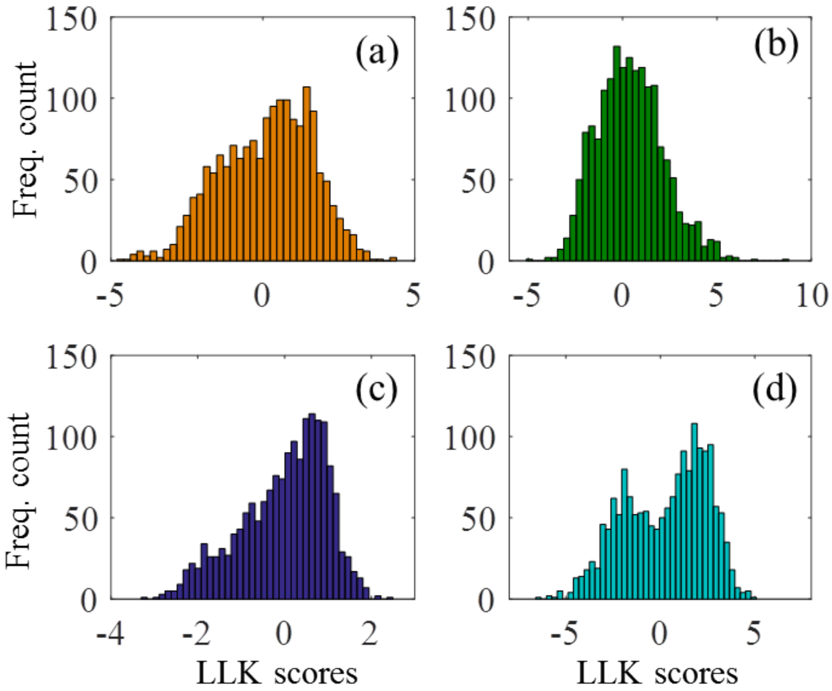


Figure 5.13: Histogram Plots of Scores of (a) CQCC, (b) LFCC, (c) MFCC, and (d) VTECC Feature Set on Development Set. After [16].

5.5.2 Results with Score-Level Fusion

In addition to the individual performance of the feature sets, we further performed the score-level fusion in order to investigate the possible complementary information of the feature sets, and reduce the % EER further. The comparison of all the feature sets along with their score-level fusion of VTECC feature set with

CQCC, LFCC, and MFCC is shown in Table 5.9. It can be observed that the individual performances on development and evaluation set has higher % EER compared to the VTECC feature set. The % EER is further reduced when the CQCC and VTECC feature sets are fused at the score-level reducing the % EER to 5.85 % and 10.94 % for fusion factor α=0.6 and α=0.8 on development and evaluation set, respectively.

The performance is also shown in Figure 5.14 with DET curves for all the feature sets along with their best score-level fusion on development and evaluation set, respectively. From Figure 5.14(a), it can be observed that for MFCC, CQCC, and LFCC show high miss probability and false alarm probability which is not a good case for the voice biometric system. However, the VTECC feature set along with score-level fusion with CQCC and MFCC feature set show the reduced miss probability and false alarm probability compared to the other feature sets. On the other hand, for evaluation set, the DET curves for all the feature sets have high probability with high false alarm rate this show that the evaluation set is challenging for given SSD taask.

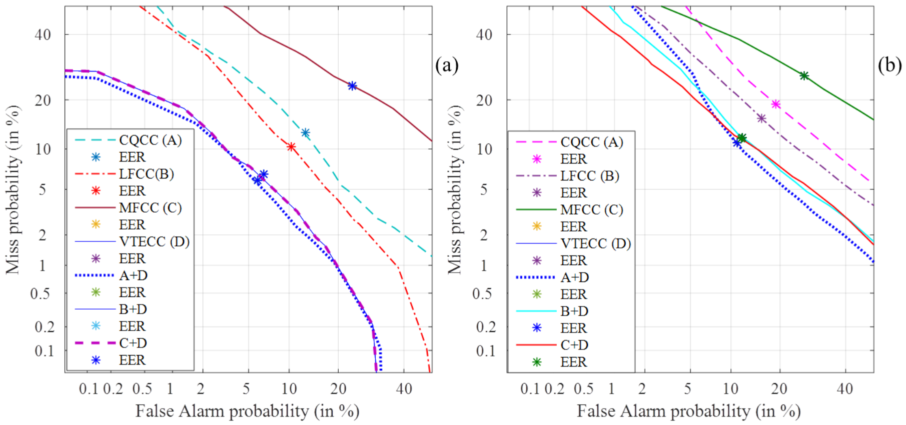


Figure 5.14: Individual DET Curves of Different Feature Sets (a) Development, and (b) Evaluation Set. After [16].

5.5.3 Results on Replay Configurations (RC)

The physical significance in terms of temporal modulations at different time scale is analyzed in Figure 5.15. The time-domain subband filtered signal around 1st formant frequency is shown in Figure 5.15(Panel I) for (a) natural, and replay with (b) perfect, (c) high, and (d) low quality devices. The slow temporal modulations of a speech signal roughly correlates with the different syllabic segments. For natural speech, slow temporal modulations results in smooth amplitude envelope as

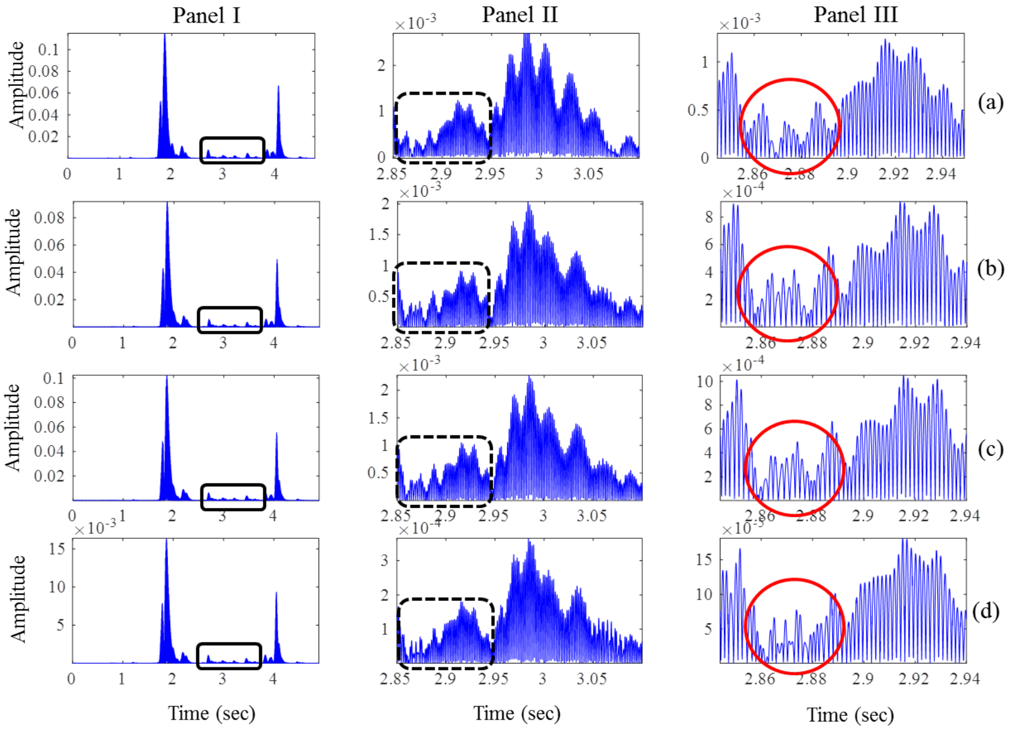


Figure 5.15: Panel I: Bandpass Filtered Signal Around First Formant. Panel II: Zoomed Version of the Above Signal Panel I. Panel III: Temporal Fine Structure of Panel II with Different Time Scale (a) Natural, and Replay with (b) Perfect, (c) High, and (d) Low Quality Devices. After [16]

shown in Figure 5.15(a) (in Panel II). The higher peaks in the fast temporal modulations (which are also known as Temporal Fine Structure (TFS)) as shown in Panel III of Figure 5.15(a) represents the harmonic structure of the speech signal. However, this observation is missing for the replay speech (Panel II) of Figure 5.15(b-d). The slow temporal modulations for replay speech are having distorted amplitude envelope (Panel II) of Figure 5.15(b). While the fast temporal modulations do not represents the harmonic structure Figure 5.15(b-d) of Panel III. It can be observed from the slow temporal modulations of replay speech that the variations are very less. On the other hand, the fast temporal modulations indeed show the differences for different quality of intermediate devices varying from the perfect, high, and low. The perfect and high quality device have the similar pattern of fast temporal modulations, however, this analysis could be very useful for the speech signal when recorded in low quality devices (as observed in Panel III of Figure 5.15(d)).

The level of noise in acoustic environment, playback, and recording device are assumed to be *inversely* proportional to the threat for ASV system pose [6]. The Replay Configurations consists of acoustic environment, playback, and recording devices, respectively. These RCs are further classified into three different threat levels, namely, low, medium, and high. Different environments have the variations with the levels of additive ambient, convolutive, and reverberation noise. According to the different RC, the % EER of VTECC feature set along with CQCC, LFCC, and MFCC are shown in Figure 5.16. The least % EER for all the RCs are obtained with the proposed VTECC feature set. It can be observed that for all the RC, the % EER for MFCC feature set are too high compared to the LFCC and CQCC feature sets. The high-level threats are difficult to detect due to use of professional audio equipment, such as active studio monitors, studio headphones, etc. to produce replay samples [6].

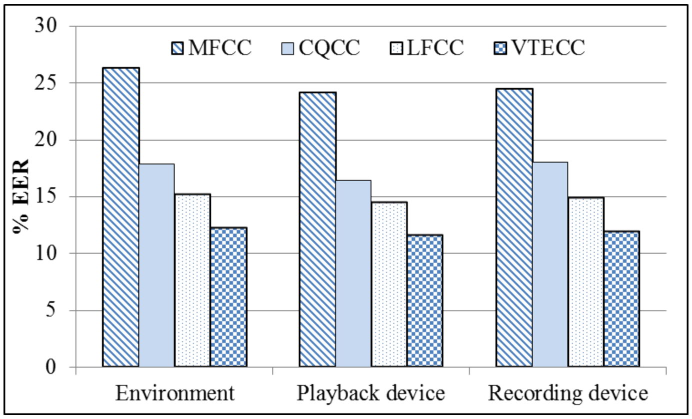


Figure 5.16: Bar Graph Representation for Different Replay Configurations, i.e., Acoustic Environment, Playback, and Recording Devices (Results in % EER). After [16].

5.6 Chapter Summary

The extension of earlier speech demodulation-based features using Variable length version of ESA, i.e., VESA is presented in this Chapter for replay SSD task, to classify the replay attack from the natural speech. We investigated the advantage of VESA over HT and ESA by varying the Dependency Index (DI) to capture the *hidden* dependencies and dynamics. The features obtained from different demodulation techniques gave better results than the baseline system of both ASVspoof 2017 challenge v2.0, and BTAS 2016 database. The limitation of the work is to investigate why particular DI gives better performance. In the next Chapter, we will present the significance of using combination IA and IF components for SSD task.

CHAPTER 6

AM-FM Features

6.1 Introduction

Earlier we discussed about the individual performance of IA and IF-based features either with ESA (Chapter 4), VESA (Chapter 5) or HT-based speech demodulation techniques. In this Chapter, we will discuss the importance of using the combined information of IA and IF components for replay SSD task in Section 6.2. The Amplitude Weighted Frequency modulation feature set is discussed in Section 6.3. The SSD experimental setup and results performed on standard dataset is discussed in Section 6.4 and Section 6.5, respectively. We also performed the experiments on different classifiers along with the analysis of different replay configurations in Section 6.6. Finally, in Section 6.7, we summarize the Chapter.

6.2 AM-FM Features

The AM-FM modulation features corresponding to the i^{th} subband are extracted from the i^{th} instantaneous frequency (IF), $f_i(t)$, and corresponding amplitude envelope, $a_i(t)$, where i=1, 2, 3,...., N, and N is the number of filtered subband signals [210], i.e.,

$$r_i(t) = a_i(t)\cos\left(2\pi\int_0^t f_i(\tau)d\tau\right), \tag{6.1}$$

where $r_i(t) \approx s(t) * g_i(t)$, $s(t)$ is the speech signal, and $g_i(t)$ is the impulse response of the i^{th} Gabor subband filter. The impulse response of the Gabor filter is given as [52]:

$$g(t) = \exp(-b^2t^2)\cos(\omega_c t), \tag{6.2}$$

where ω_c is the center frequency of the subband filter, and b is a parameter for controlling the bandwidth of a subband filter. The corresponding discrete-time impulse response, $g(n)$, is the sampled version of Eq. (6.2). Gabor representation is known to have *optimal* joint time-frequency resolution (i.e., Heisenberg's uncertainty principle in signal processing framework [170] (for more details, please refer Chapter 3, Section 3.4)). During feature extraction, choice of a particular filterbank structure, and demodulation technique affects the performance of SSD system.

The placing and bandwidth of the subband filter depends upon the problem of study at hand. As in the earlier studies, it is stated that in case of speaker discrimination, the linear scale performs well whereas for Automatic Speech Recognition (ASR) problem, Mel scale has shown the better performance than the other frequency scales [211], [143].

In this Chapter, Energy Separation Algorithm (ESA) is used as one of the signal demodulation technique among other approaches. According to the AM-FM model, the formant frequencies are not constant during a single pitch period (T_0=$1/F_0$), rather they vary around the center frequency that is approximated as IF in a particular subband. The study of auditory neurons indicates that the analysis of the joint AM-FM signals have more spectral information than analyzing AM and FM alone via a demodulation technique [212]. These small variations around the center frequency is captured by the Weighted Mean Frequency (WMF) denoted as F_i [210]:

$$WMFCC = F_i = \frac{\sum_{n=0}^{L} a_i^2[n]\Omega_i[n]}{\sum_{n=0}^{L} a_i^2[n]}, \tag{6.3}$$

$$WACC = \sum_{n=0}^{L} a_i^2[n], \tag{6.4}$$

$$AWFCC = \sum_{n=0}^{L} a_i^2[n]\Omega_i[n], \tag{6.5}$$

where L is the time window length, $a_i[n]$ and $\Omega_i[n]$ are estimated with Eq. (4.3), and Eq. (4.4). The instantaneous signals $a_i^2[n]$ are used as a weighting signal for the estimation of the F_i. The studies reported in [197,213] shows that the (squared amplitude) weighted F_i in time-domain is equivalent to the first and second central spectral moments of the signal, and explains weighted estimates are more robust than the unweighted ones. The weighted frequency estimate F_i provides more accurate formant frequencies, and is more robust to low energy or noisy frequency bands [206]. The AM and/or FM also capture additional acoustic information pertaining to speech production mechanism, such as nonlinear source-filter interaction, energy transfer, mode-locking behavior due to non-linearities in speech production mechanism, etc. [206].

6.3 Proposed Feature Extraction Algorithm

The block diagram of the AM-FM-based feature extraction process is shown in Fig. 6.1. The study reported in [135] states that the higher frequency regions play a major role to discriminate the replay *vs.* natural speech signal. To obtain a narrowband speech signal, we used $N = 40$ linearly-spaced Gabor filterbank with each subband filters having 200 Hz bandwidth. These extracted narrowband filtered signals are demodulated using ESA to get IA ($a_i[n]$) and IF ($\Omega_i[n]$). We extracted the features for Weighted Mean Frequency (WMF) (F_i) for each filtered signal as per Eq. (6.3). Furthermore, these F_i's were averaged over a short-time window of 20 ms, and 10 ms window shift to obtain L-dimensional Weighted Mean Frequency Coefficients (WMFC) for each frame. The low-dimensional feature vectors are obtained by applying Discrete Cosine Transform (DCT) (that has energy compaction property [214]) on WMFC, and will be denoted as Weighted Mean Frequency Cosine Coefficients (WMFCC). We have used post-processing Cepstral Mean Normalization (CMN) technique, to match the distributions of the signal.

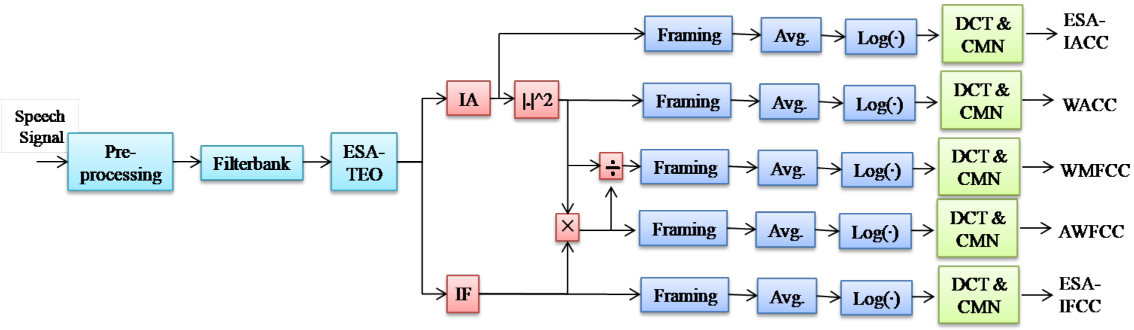


Figure 6.1: Block Diagram of Proposed Feature Sets. After [18, 29].

A similar process was followed to extract Amplitude Weighted Frequency Cosine Coefficients (AWFCC) [17]. The AWFCC feature set is the combination of IF, and squared amplitude obtained for i^{th} subband signal. This combination of IF and squared amplitude is averaged over a short-time window of 20 ms with a shift of 10 ms to obtain L-dimensional Amplitude Weighted Frequency Coefficients (AWFC) for each frame followed by DCT and CMN. Similar procedure was used to obtain Weighted Amplitude Cosine Coefficients (WACC) that contains *only* denominator term in RHS of Eq. (6.3).

6.3.1 Spectrographic Analysis

The spectral energy densities obtained from the Constant Q Transform (CQT) and the Amplitude Weighted Frequency modulation features are shown in Fig. 6.2(b) and Fig. 6.2(c) for a time-domain natural speech signal (Fig. 6.2(a)). The differences are shown by the box. In particular, the energy obtained by the constant Q transform shows the distorted formants and harmonics in lower as well as in higher frequency regions, which is captured better by the proposed feature set.

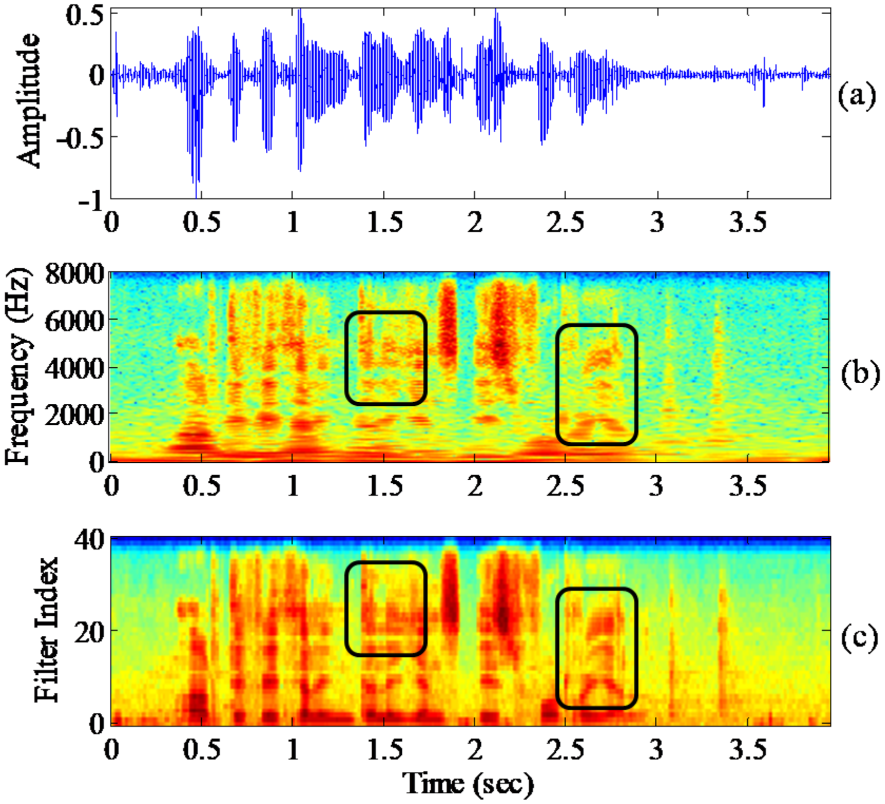


Figure 6.2: (a) Natural Speech Signal, and Corresponding Spectral Energy Density of Constant Q Transform, and Amplitude Weighted Frequency Modulation Feature Extracted Using 40 Subband Filtered Signals. The Differences are Clearly Visible in the Higher Frequency Regions (Highlighted by the Box).After [18].

The difference in spectral energy density of the AM-FM modulation feature sets is shown for natural speech signal in Fig. 6.3(a). We have compared the spectral energy density of amplitude weighted frequency (AWF) features (numerator of RHS of the Eq. (6.3)) in Fig. 6.3 (b) along with squared weighted amplitude (WA) features in Fig. 6.3 (c) and at last, weighted mean frequency (WMF) features in Fig. 6.3 (d). The differences in spectral-domain are highlighted with the circle and box. The spectral information shown in Fig. 6.3 (b) captures both amplitude and frequency information of a speech signal. On the other hand, the spectral

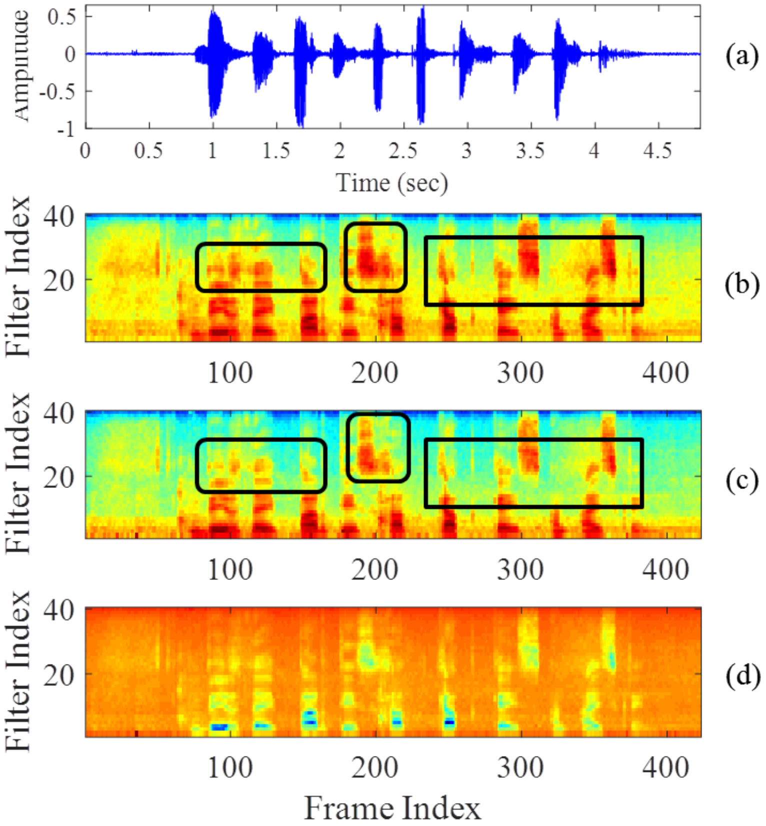


Figure 6.3: (a) Natural Speech Signal and Corresponding Spectral Energy Density Using 40 Subband Filtered Signals of Natural Speech Signal of (b) AWF, (c) WA, (d) WMF Modulation Features. The Differences are Clearly Visible in the Higher Frequency Regions (Highlighted by the Circles and Box). After [18].

information of Fig. 6.3 (c) (only squared weighted amplitude), and Fig. 6.3 (d) (weighted mean frequency) do not have much of the spectral energy information. The dynamic range of the Fig. 6.3 (d) is different as the energy present is very low as per Eq. (6.3).

The similar analysis is observed for replay speech signals recorded in different acoustic environments, namely, Panel I: Office, Panel II: Balcony, Panel III: Bedroom, and Panel IV: Canteen (as shown in Fig. 6.4). It can be observed from Fig. 6.4 that the proposed approach of using amplitude weighted frequency components (numerator of Eq. (6.3)) captures discriminatory information, and this may be useful for SSD task. In particular, the frequency regions that are affected by bandpass characteristics of replay device, mic, loudspeaker, etc. gets emphasized via stronger spectral energy density due to the term, $\sum_{i=0}^{L} a_i^2[n]\Omega_i[n]$.

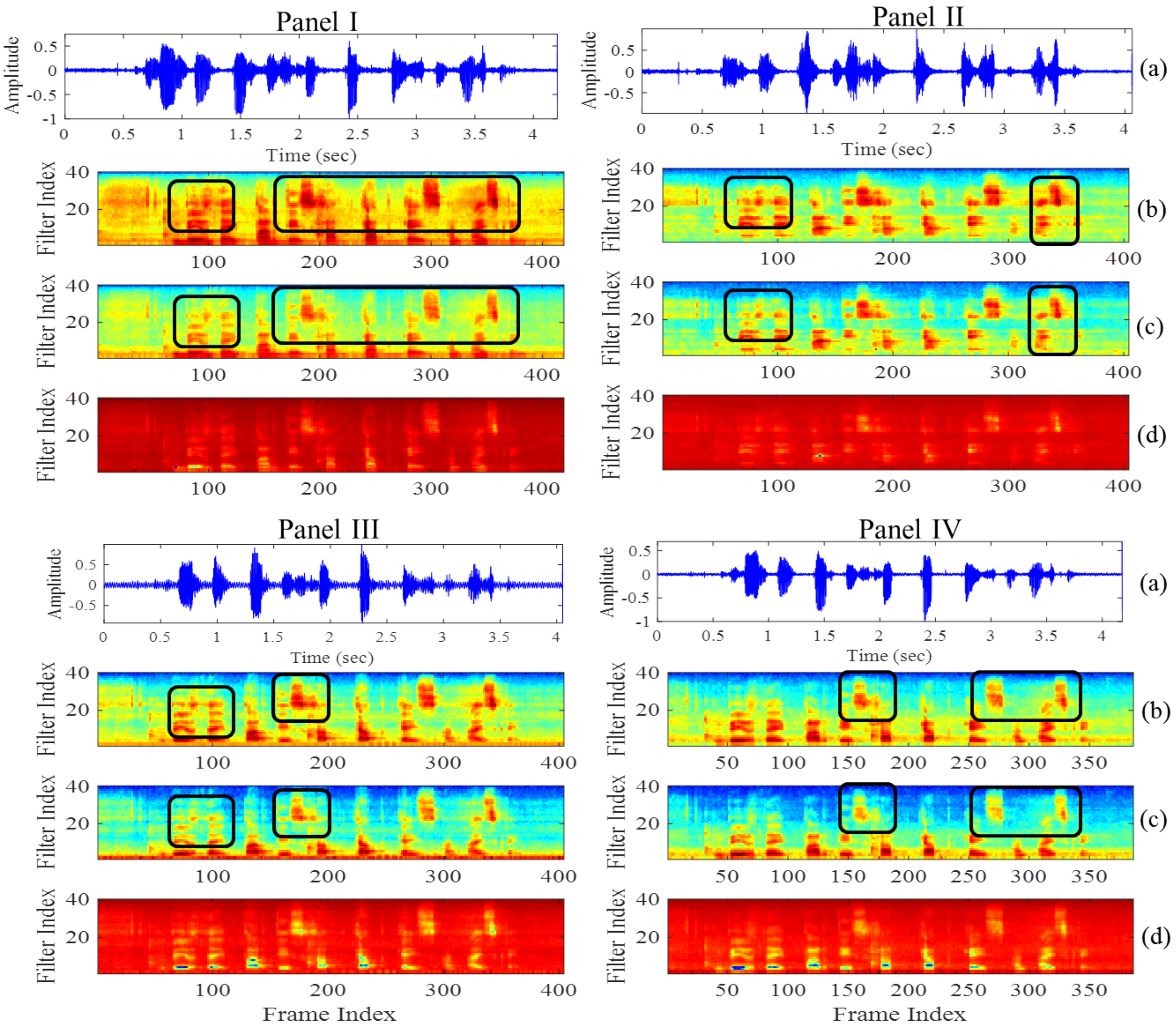


Figure 6.4: (a) Replay Speech Signal and Corresponding Spectral Energy Densities of (b) AWF (c) WA (d) WMF Modulation Features Using 40 Number of Subband Filtered Signals for Replay Speech Signals Recorded in Panel I: Office, Panel II: Balcony, Panel III: Bedroom, and Panel IV: Canteen (Highlighted Regions Shows the Spectral Energy Difference Between Different AM-FM Modulation Features). After [18].

6.4 Experimental Setup

Since this Chapter is an extension of our recent study reported in [29], we showed the results on ASVspoof 2017 version 1.0 database as the modified version of database was yet to be released during that time. To bring a continuation to this study with our earlier study, we have to keep results for both the versions of ASVspoof 2017 challenge databases. Hence, for comparison of experimental results, we used both ASVspoof 2017 challenge version 1.0 and 2.0 database, which is mainly based on the RedDots corpus, and its replay speech [4–6, 105]. The version 2.0 database presents in depth analysis of the replay detection performance along with description of playback, and recording devices. Furthermore, ASVspoof 2017 challenge version 2.0 database was released to correct data anomalies that were detected in the post evaluation of version 1.0 database. The spoofed data was recorded through different acoustic environments in the H2020-funded OCTAVE project [104]. The details of both version 1.0 and 2.0 databases are given in Table 6.1 [5].

Table 6.1: Statistics of ASVspoof 2017 Challenge Corpus. After [5, 6]

Subset	# Speakers	ASVspoof Version 1.0		ASVspoof Version 2.0	
		Genuine	Spoofed	Genuine	Spoofed
Training	10	1508	1508	1507	1507
Development	8	760	950	760	950
Evaluation	24	1298	12008	1298	12008

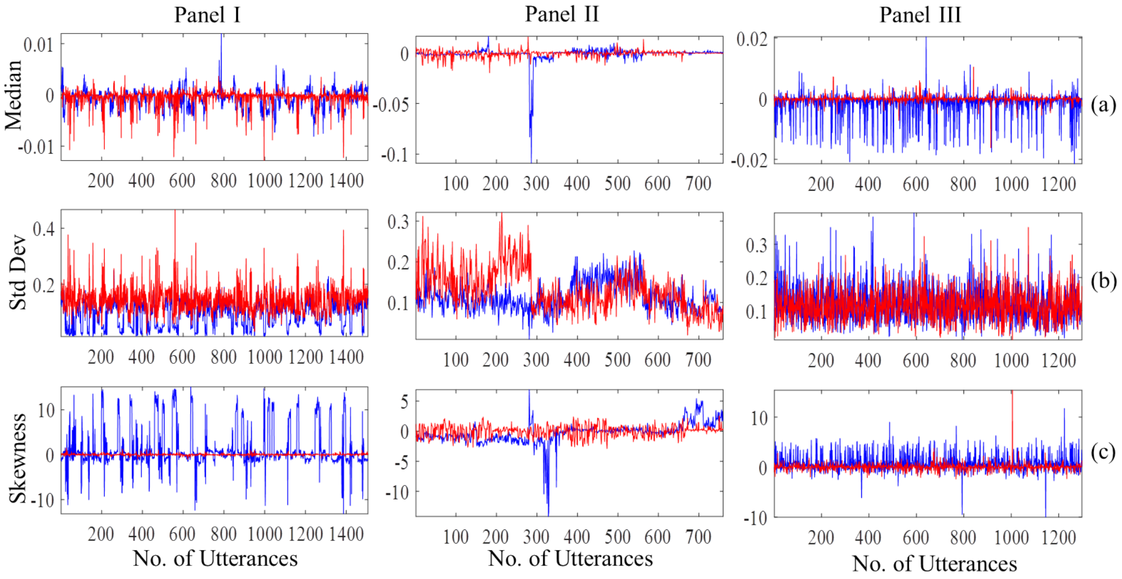


Figure 6.5: Comparison of Sample Statistics of (a) Median, (b) Standard Deviation, and (c) Skewness for Natural (Blue Line) and Replay (Red Line) Utterance. Panel I: Training, Panel II: Development, and Panel III: Evaluation Set of ASVspoof 2017 Challenge Version 2.0 Database. After [18].

Furthermore, we estimated the sample statistics, such as median, standard deviation, and skewness for every utterance from the ASVspoof 2017 challenge version 2.0 database [6]. The plots of these sample statistics are shown in Fig. 6.5 for training (Panel I), development (Panel II), and evaluation (Panel III) datasets. The values of sample statistics for each utterance corresponding to natural speech are plotted with blue color and the replay speech are shown with the red color. From these plots, it can be observed that the replay speech do not contain the same sample statistics as that for natural speech. This indeed shows that the speech samples of both natural and replay signals are *statistically* different. In particular, training and development datasets, the sample statistics of median do not have much difference for natural and replay signals, whereas for evaluation set, we can see the difference (shown in Fig. 6.5(a)). The sample statistics obtained from standard deviation, and skewness shows the difference for all the datasets (Fig. 6.5 (b and c)). The standard deviation of training and development have more difference between natural and replay speech signal, whereas for evaluation set, they overlap with each other showing less difference with each other and hence, possibly create difficulty for SSD task.

Following state-of-the-art feature sets were used for comparison with proposed feature set for replay SSD task. Constant Q Cepstral Coefficients (CQCC) features are extracted with the CQT [112]. Features are extracted with 30 DCT static coefficients appended with Δ and $\Delta\Delta$ resulting in *90*-D feature vector. The Linear Frequency Cepstral Coefficients (LFCC) features are extracted from linearly-spaced triangular filterbank energies followed by the DCT to obtain cepstral coefficients using 20 ms window with a 10 ms shift. Features extracted with 40 subband filters in a filterbank resulting in *120*-D (static+Δ+ $\Delta\Delta$). Mel Frequency Cepstral Coefficients (MFCC) features are extracted by performing STFT analysis over Hamming windowed segments of speech with 25 ms duration, along with a shift of 10 ms. The resultant power spectrum is warped using a filterbank having 40 subband filters to give 13 DCT static coefficients appended with their Δ and $\Delta\Delta$ resulting in *39*-D feature vector.

The proposed feature sets were extracted from 40 linearly-spaced Gabor filterbank with 40 DCT static and 40-Δ and 40-$\Delta\Delta$ coefficients resulting in total *120*-D feature vector. We have used Gaussian Mixture Model (GMM) classifier with 512 number of mixtures for modeling the classes corresponding to natural and replayed speech utterances. Final scores are represented in terms of Log-Likelihood Ratio (LLR). The decision of the test speech being natural or spoofed is based on the scores of LLR:

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \tag{6.6}$$

where $P(X|H_0)$, and $P(X|H_1)$ are the likelihood scores of natural and replay trials (with hypothesis H_0 and H_1), respectively. The score-level fusion is given by:

$$LLR_{fused} = \alpha LLR_{feature1} + (1-\alpha)LLR_{feature2}, \tag{6.7}$$

where $LLR_{feature1}$ is a log-likelihood score of CQCC or LFCC or MFCC, and $LLR_{feature2}$ represents score for the proposed feature set. The fusion parameter, α lies between $0 < \alpha < 1$ to decide the relative weight of scores.

6.5 Experimental Results

In this Section, we have performed experiments on both the ASVspoof 2017 challenge version 1.0 [5], and version 2.0 databases [6]. In the first set of experiments, the performance was obtained with analyzing the effect of different subband filtered signals. Furthermore, we performed experiments on evaluation dataset of ASVspoof 2017 version 2.0 database with different replay configurations. In addition, we also observed the performance on the different levels of threats, namely, low, medium, and high for replay configurations.

6.5.1 Results of Various AM-FM Features

The spectral energy density obtained via traditional spectrogram and via AWFCC feature set is shown in Fig. 6.6(a) and Fig. 6.6(b). The Panel I and Panel II corresponds to the spectral energy density for natural and replay speech signal, respectively. From the Fig. 6.6, it can be observed that the spectral energies obtained from the AWFCC feature set gave high energies in both lower and higher frequencies as compared to the spectrogram. The similar pattern was observed for the replay speech signal (Panel II). Highlighted regions in Fig. 6.6 shows the energy differences corresponding to the natural and replay signals. These energies obtained from the proposed approach contributes to detect replay signal.

6.5.2 Frame-Level Analysis for Speech Signal

Fig. 6.7 shows the analysis of natural and replay speech signals with Log-Likelihood (LLR) scores obtained from the GMM classifier at the frame-level for CQCC (Fig. 6.7(a)), and AWFCC (Fig. 6.7(b)) feature sets. It can be observed from Fig. 6.7 that the LLR scores for natural speech (Panel I) have the scores above the threshold (in this case, the threshold is 0). The replay signals are recorded in different acoustical environments, such as balcony (Panel II), canteen (Panel III), and office (Panel IV). The AWFCC feature set perfectly detects the natural and replay signals that can be observed from the LLR scores (as shown in Table 6.2). With AWFCC feature set, the replay speech signals scores obtained are negative whereas for natural speech, the scores are positive. On the other hand, CQCC feature set fails to

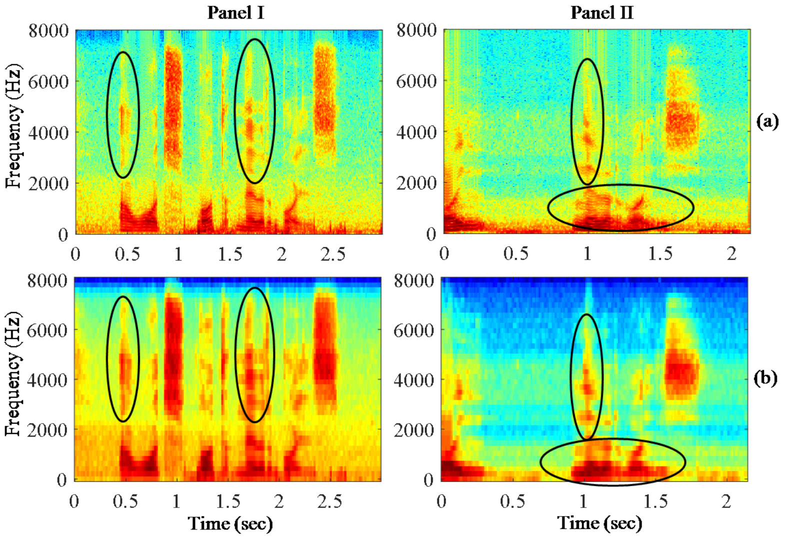


Figure 6.6: Comparison of spectral energy density via (a) spectrogram, and (b) using AWFCC feature set. Panel I and Panel II shows natural and corresponding replay speech. Highlighted oval regions shows the difference in pattern of spectral energy distribution. After [18].

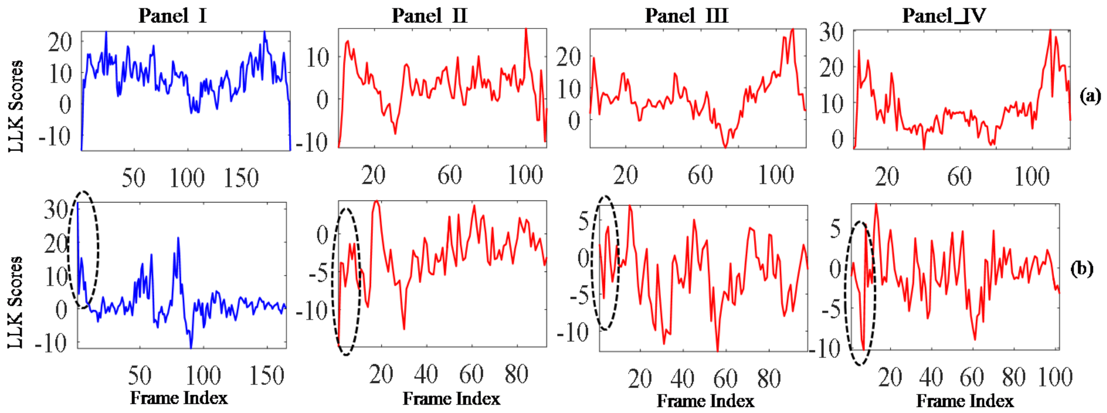


Figure 6.7: Frame-level analysis of LLR scores obtained for ASVspoof 2017 Challenge database via (a) CQCC, and (b) AWFCC feature sets. Here, Panel I: Natural, Panel II: Balcony, Panel III: Canteen, Panel IV: Office. Highlighted circle shows the difference in LLR scores for initial frames because of silence regions. After [17].

detect the speech signal that belongs to the replay class. This is analyzed with the frame-level LLR scores that are obtained from the GMM classifier. With AWFCC feature set, the initial frames of the signal shows strong peak values for natural signal that indeed helps to classify the signal that belongs to the natural class. On the other hand, for all the replay signals, the initial frames of the signals do not show these strong peaks and the scores are negative and hence, they detect the speech signal in replay class. This observation was not found with CQCC feature set, for replay speech signals, the LLR scores are positive and hence, fails to detect the speech signal in replay class. This finding is in line with recent study reported in [215].

Table 6.2: Scores Obtained from GMM Classifier on ASVspoof 2017 Version 1.0 Database for CQCC, and AWFCC Features Set. After [5]

Panel	Wave ID	Key	LLR Scores	
			CQCC	AWFCC
I	D_1000009	Genuine	8.6157	1.8751
II	D_1001644	Balcony	3.0774	−3.0448
III	D_1001074	Canteen	7.0273	−2.1653
IV	D_1000789	Office	8.2764	−0.7387

6.6 Classifiers Used

6.6.1 Gaussian Mixture Model (GMM)

We have used GMM classifier with 512 mixtures in order to obtain the training models for classification of the natural and replayed speech. Final scores are represented by Log-Likelihood Ratio (LLR) [216]. To obtain possible complementary information of the AWFCC feature set, score-level fusion was performed with CQCC, MFCC, and LFCC feature sets as per given Eq. (6.7).

6.6.2 Convolutional Neural Network (CNN)

The CNN architecture used is the same as proposed in [150]. This architecture consists of three convolutional layers followed by a max-pooling layer, and three Fully-Connected (FC) layers. The convolutional layer take a 2-D image of size $d \times N$ as input, where d represents the dimension of the input, and N represents the number of frames $(N{=}256)$. In this architecture, the first three convolutional layers have a filter/kernel size of [d×3, 1×3, 1×3] dimension, respectively. Each convolutional layer has *128* subband filters. The fourth layer is a max-pooling layer used with 1×2 kernel, and stride on the output of the third convolutional layer. At last, three FC layers with *256* units (neurons) and a softmax layer is used for computing the final scores. We have used dropout of 0.5 as a regularization to all the three FC layers. All the layers consists of Rectified Linear Unit (ReLU) as an activation function.

6.6.3 Long Short-Term Memory (LSTM)

The third classifier used here is LSTM. The number of neurons in the input layer is equal to the dimension of a feature vector, whereas the number of neurons in the output (softmax) layer is two (one for natural class, and another for spoofed class). The LSTM layer contains the 256 neurons. We have used the *tanh* non-linearity for each neuron. For training the CNN and LSTM models, Adam optimizer, and 90 % overlapping input data is used, which makes the networks to learn the data-dependency. We trained both the models for 100 epochs with 64 batch size.

6.6.4 Results with GMM, CNN, and LSTM Classifiers

The results obtained using AWFCC feature set are compared with CQCC, MFCC, and LFCC feature sets in Table 6.3. We have used GMM, CNN, and LSTM classifiers to obtain the corresponding models for all the feature sets. The baseline system gave an EER of 10.35 % on development (dev) set, and 28.48 % on evaluation (eval) set. The CQCC with CNN and LSTM classifiers gave an EER of 21.34 % and 22.72 % on eval set, respectively. The AWFCC feature set with GMM classifier obtained a lower EER of 6.52 % (dev) and 11.83 % (eval), while with CNN classifier, the results were comparable to the GMM classifier with EER of 6.97 % (dev) and 13.71 % (eval) set.

Table 6.3: Results in (% EER) with GMM and CNN and LSTM Classifier. After [17]

	Classifier					
Feature	**GMM**		**CNN**		**LSTM**	
Set	Dev	Eval	Dev	Eval	Dev	Eval
CQCC	10.35	28.48	11.71	21.34	**09.60**	**22.72**
MFCC	11.21	31.30	15.13	26.96	30.13	41.44
LFCC	10.58	16.62	11.84	19.33	34.73	38.21
AWFCC	**06.52**	**11.83**	**06.97**	**13.71**	27.10	33.85

However, the AWFCC feature set does not perform better with LSTM classifier as it increased the EER to 27.10 % and 33.85 % on dev and eval sets, respectively. The frame-level GMM classifier performed better due to the initial silence regions of natural speech signal, which is absent for replay signals (observed from the Fig. 6.7) [215]. These initial silence frames gave high LLR scores and hence, detects the natural signal corresponding to its replay counterparts. While, CNN and LSTM classifiers are modeled at utterance-level and hence, the effect of silence region is averaged. Hence, utterance-level classifier fails to contribute the information that relates to the silence frames and thus, degrades the performance of SSD task. In our recent study reported in [58], we observed that because of the fast fluctuations for the IF components, the natural and replay models are not trained properly for neural network-based classifiers and hence, CNN and LSTM classifiers do not perform well to detect the replay speech signals. The LSTM classifier has the property that it deals with the dependencies of the context, because of these fluctuations in IF components, it fails to capture the dependencies of the speech signal with AWFCC feature set and hence, did not perform well.

Table 6.4: Results (% EER) with Classifier-level Fusion for a Fixed Feature Set. After [17]

	Classifier					
Feature	**GMM+CNN**		**GMM+LSTM**		**LSTM+CNN**	
Set	Dev	Eval	Dev	Eval	Dev	Eval
CQCC	05.09	20.21	**04.06**	19.75	09.24	22.59
MFCC	06.83	23.83	10.36	30.54	14.10	26.49
LFCC	06.29	09.98	09.53	15.84	11.64	18.86
AWFCC	**01.78**	**06.47**	04.62	**10.77**	**06.67**	**13.60**

+ : Classifier-level fusion for a fixed feature set

6.6.5 Results with Classifier-Level Fusion

We have fused two systems at classifier-level in order to obtain the possible complementary information present in the different classifiers and the results of this kind of fusion for all the feature sets are shown in Table 6.4. It can be observed from Table 6.4 that the reduction in EER from the individual EER for each feature set. The CQCC obtained a reduced EER to 5.09 % and 20.21 % on dev and eval sets, respectively, with the fusion of GMM and CNN classifiers, and gave low EER with the fusion of GMM and LSTM classifiers resulting in reduced EER of 4.06 % and 19.75 % on dev and eval sets, respectively. The AWFCC feature set gave slightly lower % EER using classifier-level fusion for all the three classifiers considered in this study. The fusion of GMM and CNN classifiers with AWFCC feature set gave a lower EER of 1.78 % and 6.47 % on dev and eval sets, respectively, while fusion of GMM and LSTM classifiers gave EER 4.62 % and 10.77 % on dev and eval sets, respectively. On the other hand, the fusion of LSTM and CNN reduced the EER to 6.67 % and 13.60 % (than the individual classifier) on dev and eval sets, respectively. From Table 6.4, we observed that the combined system of GMM and CNN classifiers captures relatively better complementary information than the other combined systems.

6.6.6 Results with Score-Level Fusion of Feature Sets

In addition, the proposed feature set was fused at a score-level with CQCC, MFCC, and LFCC feature sets in order to investigate the possible complementary information, and results are shown in Table 6.5. We observed a reduction in EER than the individual EER feature sets.

Table 6.5: Results (% EER) with Score-Level Fusion for a Fixed Classifiers. After [17]

Classifier	Feature Sets					
	A⊕CQCC		**A⊕MFCC**		**A⊕LFCC**	
	Dev	Eval	Dev	Eval	Dev	Eval
GMM	**03.39**	**10.85**	**04.01**	**11.83**	**06.52**	**11.83**
CNN	04.32	13.42	05.83	13.33	06.96	13.60
LSTM	09.02	21.74	23.03	33.37	26.78	33.08

A: AWFCC, ⊕ : Score-level fusion of feature sets for a fixed classifier

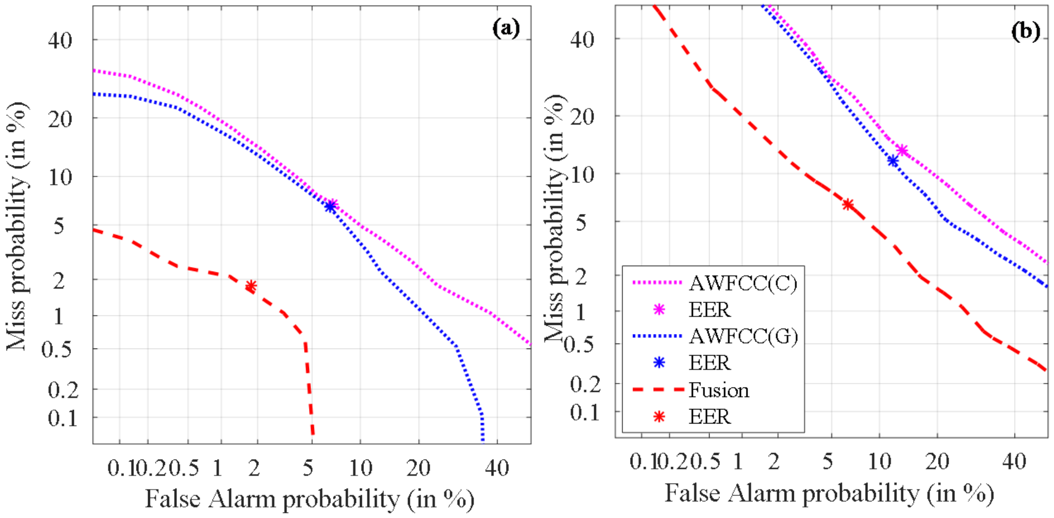


Figure 6.8: EER (%) Plots of Different Feature Sets with Different Classifiers on (a) Development, and (b) Evaluation Set. After [18].

Table 6.6: Final Results on Dev and Eval Set. After [17]

Feature Set	Classifier	Dev	Eval
CQCC (Baseline)	GMM	10.21	28.48
LFCC	GMM	10.58	16.62
MFCC	GMM	11.21	31.30
AWFCC	GMM	06.52	11.83
AWFCC+CQCC	GMM	03.39	10.85
AWFCC	GMM+CNN	**01.78**	**06.47**

The lower EER obtained is with the fusion of AWFCC and CQCC feature set with GMM classifier reducing the EER to 3.39 % on dev set (with a fusion factor, α=0.6), and 10.85 % on eval set (with a fusion factor, α=0.8). The final results using all the feature sets are shown in Table 6.6. The lower % EER is obtained with AWFCC feature set by the classifier-level fusion of GMM and CNN. The performance is also shown by the DET curves in Fig. 6.8(a) for dev set and Fig. 6.8(b) for eval set with AWFCC feature set obtained with GMM and CNN along with its scores of classifier-level fusion [217]. From the DET curves, it can be observed that both the miss probability and false alarm probability are less for the scores obtained from the classifier-level fusion, which is a good case for ASV system so far as its performance is considered.

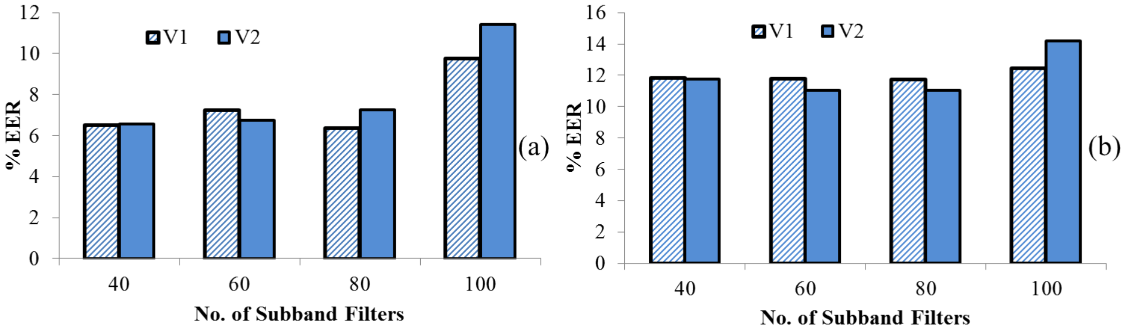


Figure 6.9: Results in % EER Showing the Effect of Number of Subband Filters on Both ASVspoof 2017 Version 1.0 (V1 Line Pattern Fill), and Version 2.0 (V2 Solid Fill) Database (a) Results on Development Set, and (b) Results on Evaluation Set. After [18].

The results of various AM-FM feature sets are shown in Table 6.7. All the feature sets were extracted using 40 number of subband filters with *120*-D feature vector that includes 40 static appended with 40-Δ and 40-ΔΔ features. The results obtained with ASVspoof 2017 challenge version 1.0 database shows the lower % EER with AWFCC feature set resulting in % EER of 6.52 % and 11.83 % on development and evaluation datasets, respectively. The AWFCC feature set indeed helps to capture the discriminatory information (in both amplitude and frequency) between natural *vs.* replay speech. The AWFCC feature set preserves both the squared amplitude information as well as the corresponding IF information. This may be the reason for the lower EER using AWFCC feature set on development and evaluation datasets than the other feature sets. When the results were compared with WMFCC feature set, it did not perform well because of lower spectral energy density, and less dynamic range. When the features were computed for AWFCC, it gave comparatively better results than the WMFCC feature sets. The similar observation is observed on the ASVspoof 2017 version 2.0 database resulting in lower % EER with AWFCC feature set compared to the other feature sets, namely, WACC and WMFCC (as shown in Table 6.7). In particular, the % EER obtained on development and evaluation sets is 6.56 % and 11.78 %, respectively.

Table 6.7: Results Using 40 Subband Filters for Various AM-FM Feature Sets. After [18]

	ASVspoof V1.0		**ASVspoof V2.0**	
Feature Set	**Dev**	**Eval**	**Dev**	**Eval**
ESA-IACC	6.48	12.00	7.99	13.45
ESA-IFCC	**4.12**	12.79	11.84	12.93
WMFCC	28.07	31.59	26.95	27.18
WACC	7.08	12.87	7.86	12.78
AWFCC	**6.52**	**11.83**	**6.56**	**11.78**

6.6.7 Results with Varying Number of Subband Filters

In the earlier Section 6.5.1, we found that the feature set that is extracted by combination of both amplitude and frequency components (i.e., AWFCC feature set) gave relatively lower % EER and hence, further set of experiments are performed for only AWFCC feature set. In this Section, we performed experiments with various number of subband filters for proposed AWFCC feature set on both the versions (1.0 and 2.0) of ASVspoof 2017 challenge database. The AM-FM features based on ESA demodulation technique are known to work better when the signal is a narrowband signal [54]. Since the proposed feature sets are extracted with linearly-spaced Gabor subband filters, the frequency resolution is better captured, and it is related to the number of subband filters used. In particular, increasing the number of subband filtered signals provides a good frequency resolution and thus, captures more detailed spectral characteristics. We have varied number of subband filters of the AWFCC feature set as shown in Fig. 6.9(a) for development set, and Fig. 6.9(b) for evaluation set on both ASVspoof 2017 version 1.0 (V1) and version 2.0 (V2). Initially, the AWFCC feature set was computed using 40 subband filters in Gabor filterbank with 200 Hz bandwidth. We further increased the number of subband filters up to 100 in Gabor filterbank. On ASVspoof 2017 version 1.0 database, we obtained lower % EER with 80 subband filters resulting in EER of 6.37 % and 11.72 % on development and evaluation sets, respectively. On the other hand, on ASVspoof 2017 version 2.0 database, we obtained reduced % EER with 60 subband filters giving an % EER of 6.74 % and 11.03 % on development and evaluation sets, respectively. By increasing the number of subband filters in a filterbank, AM-FM components might get overlapped with adjacent subband filters and hence, their estimation becomes inaccurate. In this case, the number of subband filters required are in the range of 60 to 80 to cover the entire spectrum information that may help for replay SSD task.

6.6.8 Effect of Replay Configurations (RC)

The version 2.0 database of ASVspoof 2017 challenge provides a detailed description of acoustic environment, playback, and recording devices [6]. A combination of acoustic environment, playback, and recording device is referred to as Replay Configurations (RC). The RC are again classified into three different levels of threats to the ASV system w.r.t the intermediate devices used for recording and playback. The threat to the ASV system depends on the quality of the devices as well as on the distance at which the device is placed from the source speaker to record the voice. The schematic representation of different levels of threats, namely, low, medium, and high are shown in Fig. 6.10 [5]. The high-level threat is shown with red background, where the intermediate device is placed at a very small distance from the source speaker. Similarly, medium and low-level threats are shown with blue and green background, respectively, where the distance from the source speaker to the intermediate devices is large as compared to the high-level threat. Thus, level of threat is *inversely* proportional to distance of source speaker from the intermediate devices as given by Eq. (6.8), i.e.,

$$\text{Level of Threat} \propto \frac{1}{\text{Distance of Source to Device}}. \tag{6.8}$$

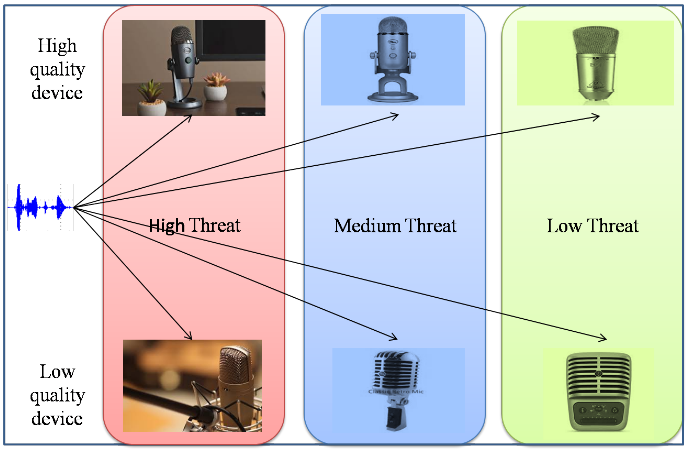


Figure 6.10: Schematic Diagram for Different Levels of Threats (Low-to-High) to the ASV System w.r.t the Quality, and Distance of the Recording Device from the Source Speaker. After [18].

6.6.8.1 Effect of Acoustic Environments

The ASVspoof 2017 challenge version 2.0 database has in total 26 different environments denoted from E01-E26 [6] . Different environments have the variations included with levels of additive ambient, convolutive, and reverberation noise. The acoustic environments were classified into three different levels of threat, namely, low, medium, and high. The individual % EER for the types of threat with all the feature sets are shown in Fig. 6.11. Form the bar graph plot, it can be observed that for low-level threat, the % EER is less compared to that of medium, and high-level threat for all the feature sets. The high-level threat is challenging to detect as the acoustic environment has any extra noise-free due to which the replay samples approximately matches to the natural speech and hence, fails to detect the replay speech signal. However, comparing performance of all the feature sets, proposed feature set produces relatively lower % EER for all the different levels of threats.

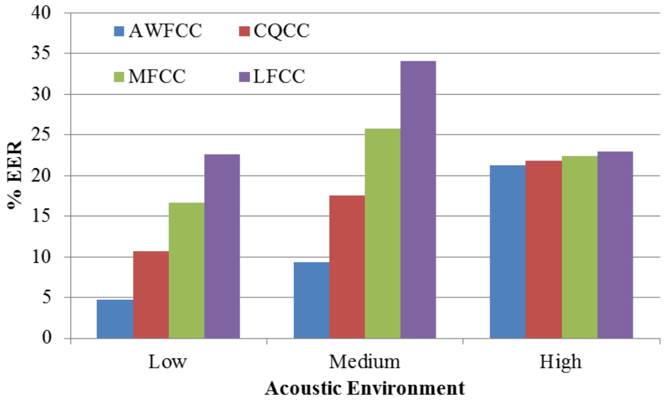


Figure 6.11: Results in % EER of AWFCC, CQCC, MFCC, and LFCC Feature Set Showing the Performance on Three Different Levels of Threats, Namely, Low, Medium, and High on Replayed Speech Signal, When Recorded in Different Acoustic Environments. After [18].

According to the levels of threat, Fig. 6.12 shows the individual performance of feature set in different acoustic environments. From the Fig. 6.12, it can be observed that for most of the cases, MFCC (green line) and CQCC (blue line) shows high % EER, whereas LFCC (violet line) and AWFCC (red line) shows lower % EER compared to the MFCC and CQCC feature sets. The proposed feature set gave relatively lower % EER compared to the other feature sets.

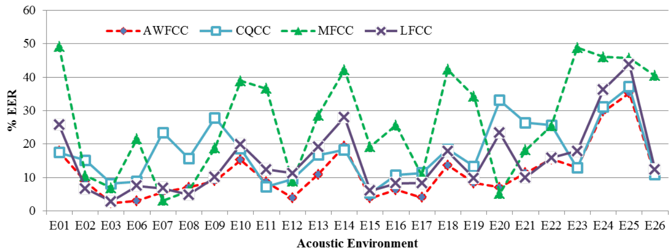


Figure 6.12: Individual % EER for Different Environment Conditions with CQCC, MFCC, LFCC, and AWFCC Feature Set. After [18].

6.6.8.2 Effect of Playback Devices

Similar to different types of acoustical environments, there are 26 different playback devices denoted by P01-P26 [6]. These playback devices also have similar levels of threats to ASV system as different acoustic environments have. The classification of different levels of threats w.r.t the feature set is shown in Fig. 6.13. It can be observed that our proposed feature set in all the levels of threats gave lower % EER than the other compared feature sets. Furthermore, for low and medium-level threats, it performed well. However, the results for high-level threats are comparable and hence, needs further investigations on high-level threat. Similar to acoustic environment, the performance (in % EER) is obtained w.r.t various playback devices (as shown in Fig. 6.14). The least % EER is obtained with the proposed feature set for all the levels of threats. The high-level threats are difficult to detect because professional audio equipment, such as active studio monitors, and studio headphones were used to record replay samples. As the level of threat increase from low-to-high, for playback devices, the % EER also increases.

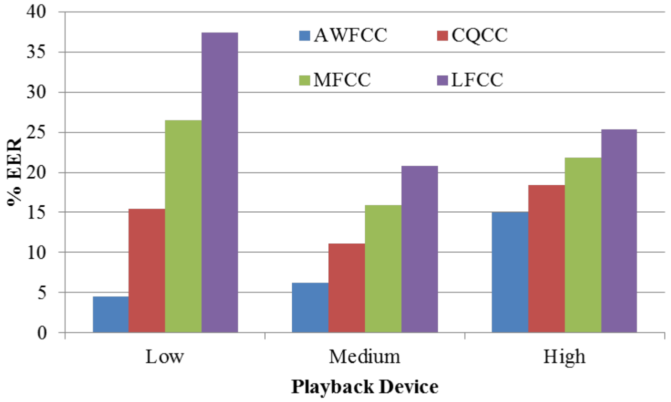


Figure 6.13: Results in % EER of AWFCC, CQCC, MFCC, and LFCC Feature Set Showing the Performance on Three Different Levels of Threats, Namely, Low, Medium, and High on Replayed Speech Signal, When Recorded in Different Playback Device. After [18].

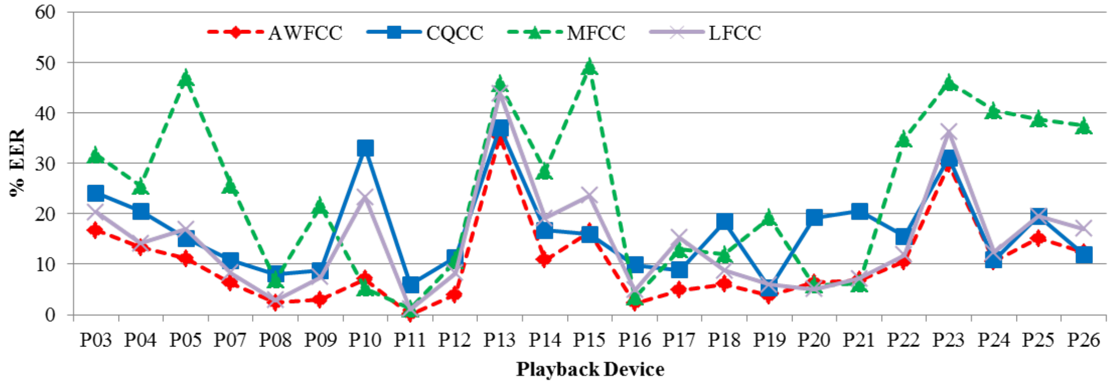


Figure 6.14: Individual % EER for Different Playback Device with CQCC, MFCC, LFCC, and Proposed AWFCC Feature Set. After [18].

Table 6.8: Results with Score-Level Fusion. After [18]

	ASVspoof v1.0		**ASVspoof v2.0**	
Feature Set	**Dev**	**Eval**	**Dev**	**Eval**
A: CQCC (90-D (SDA))	10.21	28.48	12.81	19.04
A1: CQCC (120-D (SDA))	-	-	10.99	27.26
B: LFCC	10.58	16.62	10.31	15.73
C: MFCC	11.21	31.30	24.19	26.90
D: AWFCC	6.37	11.72	6.74	11.03
A+D	**3.60**	**11.22**	**5.75**	**10.42**
B+D	6.37	11.67	6.74	11.03
C+D	3.94	11.72	6.74	10.89

+: Score-Level Fusion, SDA:Static+Delta+Acceleration

6.6.8.3 Effect of Recording Devices

There are 25 different recording devices used during collection of replay speech denoted by R01-R25 [6]. Similar to other two replay configurations, recording devices are also classified into low, medium, and high-level threats to ASV system. Similar to other two replay configurations, Fig. 6.15 shows the performance for different levels of threats to the ASV system. It can be observed that for low and medium-level threats, the proposed feature set gave lower % EER than the other feature sets with a significance difference in % EER. However, for high-level threats, very small change in % EER is observed compared to the other feature sets. The detailed results in % EER w.r.t the different recording devices are shown in Fig. 6.16 for all the feature sets, and observed that our proposed feature set relatively gave lower % EER for various recording devices considered in this study.

Furthermore, to increase the performance of the replay SSD task, we further performed score-level fusion as per Eq. (6.7) in order to explore possible com-

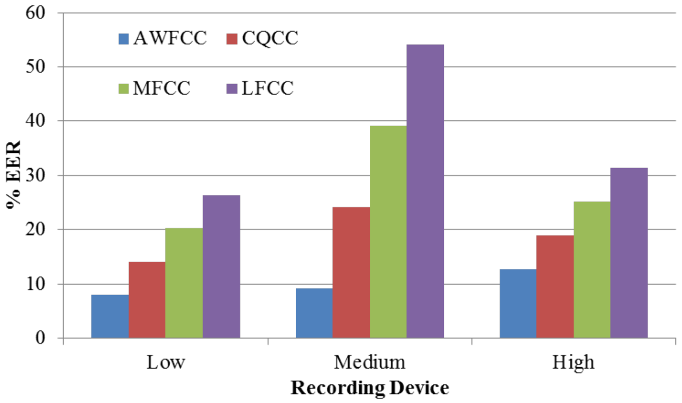


Figure 6.15: Results in % EER of AWFCC, CQCC, MFCC, and LFCC Feature Set Showing the Performance on Three Different Levels of Threats, Namely, Low, Medium, and High on Replayed Speech Signal, When Recorded in Different Recording Device. After [18].

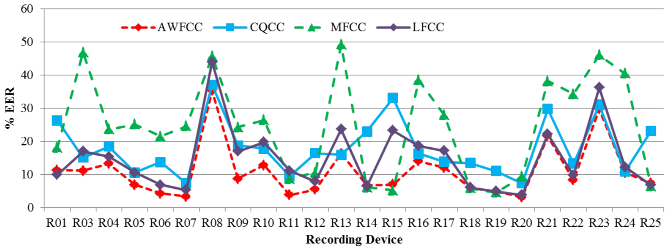
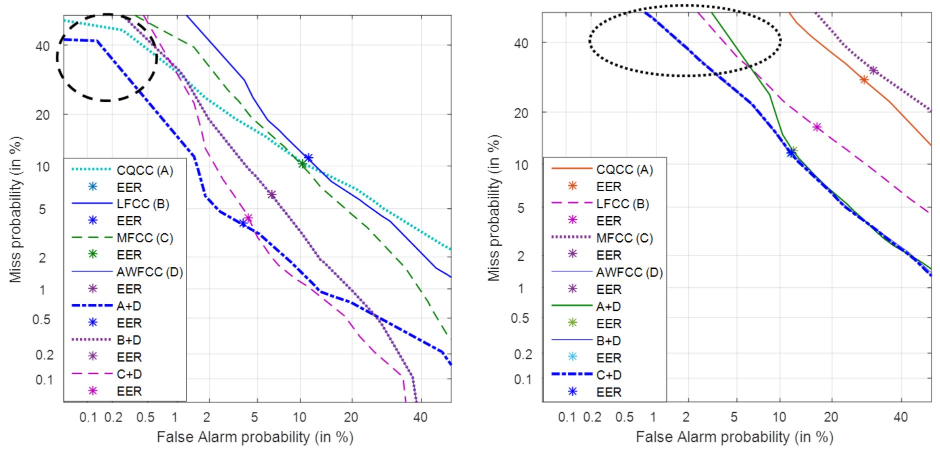


Figure 6.16: Individual % EER for Different Recording Device with CQCC, MFCC, LFCC, and AWFCC Feature Set. After [18].

plementary information. The results obtained using score-level fusion along with individual % EER are shown in Table 6.8 for both version 1.0 and 2.0 databases. It can be observed from the Table 6.8 that for both databases, the score-level fusion of AWFCC with CQCC feature set gave reduced % EER than that for CQCC alone. For ASVspoof 2017 version 1.0 database, the % EER with score-level fusion of AWFCC and CQCC gave 3.60 % and 11.22 % on development and evaluation sets, respectively. Similarly, for ASVspoof 2017 version 2.0 database, the % EER with score-level fusion of AWFCC and CQCC gave 5.75 % and 10.42 % on development and evaluation sets, respectively. This indicates that the complementary information is captured by the proposed feature set than the CQCC alone.

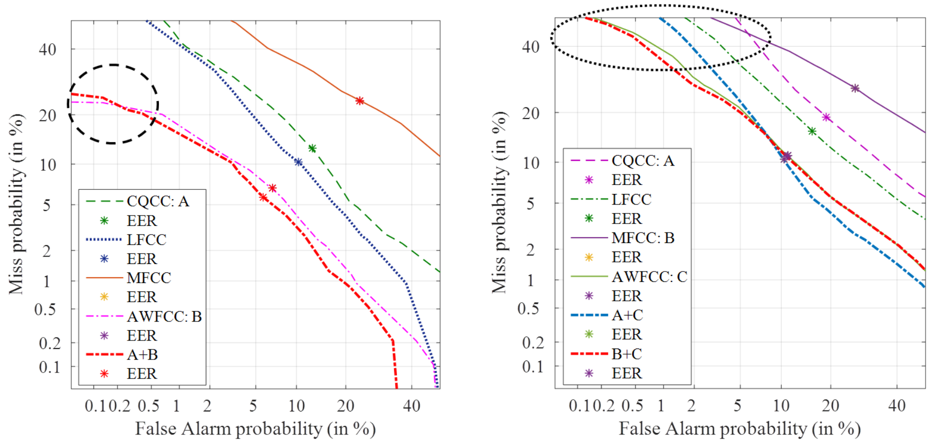
The performance evaluation is also shown by the DET curves for CQCC, MFCC, LFCC, and proposed feature set along with their best score-level fusion on both version 1.0 and 2.0 databases in Fig. 6.17 and Fig. 6.18. It is observed that for both the versions, miss probability of CQCC, MFCC, and LFCC was very high for given FAR which is not a good case for robust ASV system. There is decrease in miss probability for proposed feature set on development set as shown in Fig.

6.17(a) and Fig. 6.18(a). It can be observed from the DET curves for development set that there is reduction in the miss probability of version 2.0 DET curve (shown with dotted circle in Fig. 6.18(a)). For version 2.0, the miss probability is high for other acoustic feature sets than the proposed feature set. Similar pattern of DET curve is observed for the evaluation set on both version 1.0 and 2.0 datasets as shown in Fig. 6.17(b) and Fig. 6.18(b). The % EER is further reduced with the score-level fusion of CQCC, and AWFCC feature sets.



(a) (b)

Figure 6.17: (a) The Individual DET Curves on Dev Set with Their Score-Level Fusion, and (b) Similar DET Curves on Eval Set of ASVspoof 2017 Version 1.0 Database. Dotted Circle Shows the DET Curves with Less False Alarm Probability. After [18].



(a) (b)

Figure 6.18: (a) The Individual DET Curves on Dev Set with Their Score-Level Fusion, and (b) Similar DET Curves on Eval Set of ASVspoof 2017 Version 2.0 Database. Dotted Circle Shows the DET Curves with Less False Alarm Probability. After [18].

6.7 Chapter Summary

In this Chapter, our efforts to improve the performance of SSD task using combination of IA and IF components is presented. The IA component obtained from the Amplitude Modulation (AM) component of a narrowband speech signal is severely affected by the noise and multipath interference (due to replay mechanism). The noise present in the replayed signals are exploited via the IF components. In particular, this damage in AM components is exploited by the proposed feature set. Experiments on ASVspoof 2017 challenge v2.0 database were performed and compared the results with our earlier proposed approach and baseline system. The significance of using both the IA and IF components in the same pipeline improved the SSD performance. The limitation of the work is to investigate the individual contribution of IA and IF components for the SSD task. In the next Chapter, we will discuss about various applications, such as ASR, ASC, VCS, and WSD using the TEO-based cepstral, and spectral features on standard datasets.

CHAPTER 7
Other Applications

7.1 Introduction

In the Chapters 3-6, we discussed our proposed feature sets using ESA and TEO and applied on speech signals for the various SSD tasks (such as, SS, VC, and replay). However, our proposed feature sets is able to discriminate the variety of natural/audio sounds, such as human speech *vs.* non-speech/non-music audio samples. In this Chapter, to explore the potential of our proposed feature sets in variety of natural speech, we considered the Automatic Speech Recognition (ASR) in Section 7.2. For the ASR task, we extracted filterbank energy using TEO and abbreviated as GTFB (Gabor Teager Filterbank energy) on near and far-field speech recognition corpus, namely, LibriSpeech and CHiME-3 challenge database, respectively. Section 7.3 discusses the study on development of countermeasures for replay attacks on the Voice Assistant (VA) task on the ReMASC corpus. Furthermore, we also explored the Teager energy-based feature set for Whisper Speech Detection (WSD) task on the wTIMIT, and CHAINS corpus in Section 7.4. In addition, we also explored the proposed feature sets for the acoustic scene classification (ASC) task on DCASE 2018 challenge database discussed in Section 7.5. Furthermore, we also explored the Teager energy-based feature set for Whisper Speech Detection (WSD) task on the wTIMIT, and CHAINS corpus in Section 7.4.

7.2 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a task that converts a speech signal into a continuous sequence of words along with real interaction between humans and the machines [218]. Due to huge commercial success of Voice Assistants (VAs) or Intelligent Personal Assistants (IPAs) the far-field speech recognition is an essential technology for interactions that aims to provide access of the smart devices through the recognition of far-field speech [219]. This technology is applied to smart home appliances (smart loudspeaker and TV), meeting transcription, and onboard navigation, etc. However, in a real environment, there is a lot of background noise, multipath reflections, reverberation, and even human voice interference, leading to decrease in ASR accuracy [220].

Recent development in acoustic modelling uses approaches, such as deep learning, sequence modelling, etc., however, their performance detoriates in the case of far-field recording conditions. The reverberant artifacts distort the speech signal by smearing the amplitude envelopes of the speech signal [11]. The development of a real-world applications faces a notable challenge because of reverberation. The ASR system degrades the performance, when the far-field microphone array signals are used instead of close-talking microphone.

The aim of the 3^{rd} CHiME (Computational Hearing in Multisource Environments) challenge was to develop a multichannel ASR system [221]. The CHiME-3 dataset upgrades the difficulty by providing not only artificially noisy speech (i.e., obtained by combining clean speech with recorded background noise) but also consists of the noisy speech recorded in public environments, such as cafe, bus, street junction, and pedestrian areas. The CHiME-3 challenge covers different speakers, noise environments, and real-world problems, such as clipping, microphone failure, recording glitches, etc.

Our goal in this work is to increase the robustness of ASR using Teager energy-based features in noise and reverberation in order to combine them efficiently with standard Mel Frequency Cepstral Coefficients (MFCC)-based front-ends with GMM (Gaussian Mixture Model), and DNN (Deep Neural Network) acoustic models in addition to use of RNN (Recurrent Neural Network) as language models. The motivation behind using TEO is its attributes to capture nonlinear aspects of speech production [190]. The true total energy of source is estimated using TEO, and it also preserves the amplitude and frequency modulation of a resonant signal and hence, it improves the time-frequency resolution along with improving the formant information representation [52]. In addition, the TEO has the noise suppression property (for details, please see Appendix C), and it attempts to suppress the distortion caused by the noise signal. While there are studies in the ASR literature that exploit noise suppression capability of TEO for ASR, however, they are reported either for close-talking speech [164] or artificially added noise [222], the present study extends this work for a typical noise characteristics of real far-field scenarios.

The TEO tracks running estimate of instantaneous energy fluctuations of a

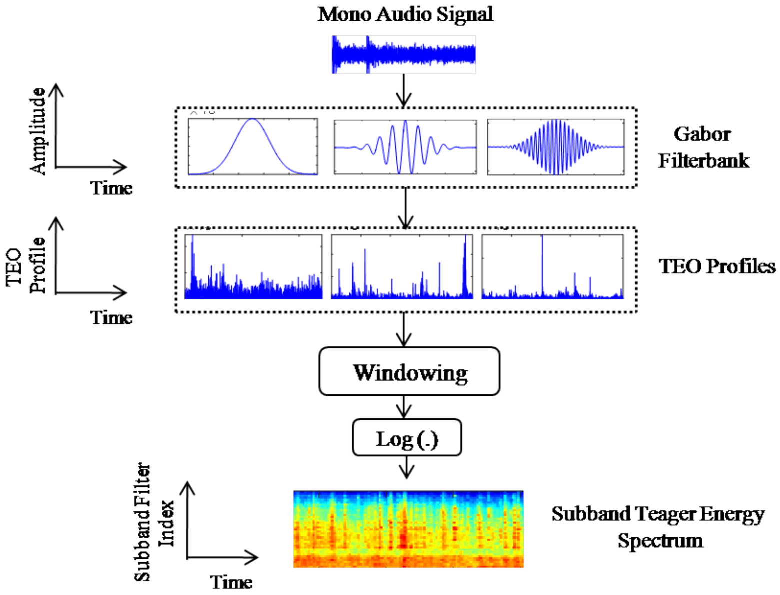


Figure 7.1: Block diagram of Teager energy spectral features. After [19, 23].

narrowband speech signal as discussed in Chapter 3, [51,52,190], i.e.,

$$\Psi_d\{x_i[n]\} = x_i^2[n] - x_i[n-1]x_i[n+1] \approx a_i[n]^2\Omega_i[n]^2, \tag{7.1}$$

where $x_i[n]$ is discrete-time bandpass filtered signal for i^{th} subband filter, and $\Psi_d\{\cdot\}$ represents TEO for discrete-time signals. As discussed earlier, the TEO works on narrowband signal and hence, bandpass filtering is necessary to apply on the input speech signal to compute *'N'* number of subband filtered signals. The block diagram of Gabor filterbank energy coefficients-based on TEO is shown in Fig. 7.1. The feature extraction procedure is similar as discussed in Chapter 3, Section 3.4.

In ASR, the lower formants and harmonics are important as the speech information is present in lower frequencies and hence, they should be preserved. Furthermore, these subband filtered signals are given to the TEO block in order to estimate the Teager energy profile of each subband filtered signals. These Teager energy profiles are further passed through the frame-blocking along with averaging of the speech segment using a window length of 25 ms and shift of 10 ms followed by logarithm operation. Finally, these filterbank energy coefficients are extracted from the speech signal. Henceforth, we will denote it as Gabor Teager Filterbank (GTFB) feature set for the ASR task.

The spectral energy densities of the speech signal from the CHiME-3 corpus is shown in Fig. 7.2. The comparison is done of the spectral energy densities obtained from the Mel filterbank, and GTFB features as shown in Fig. 7.2(b), and

Fig. 7.2(c), respectively, for both clean (Panel I) and noisy (Panel II) speech signal. It can be observed from the spectral energy densities that the energy obtained for the clean speech signal from the Mel filterbank has less energy density compared to the GTFB approach. For the ASR task, the lower formants, such as F_1 and F_2 are important to preserve the speech content. The spectral energy obtained from the TEO shows the sharp formants, and high energy compared to that of Mel filterbank features. In particular, the Mel spectral energy obtained are distorted, and have blur characteristics at the higher frequency regions.

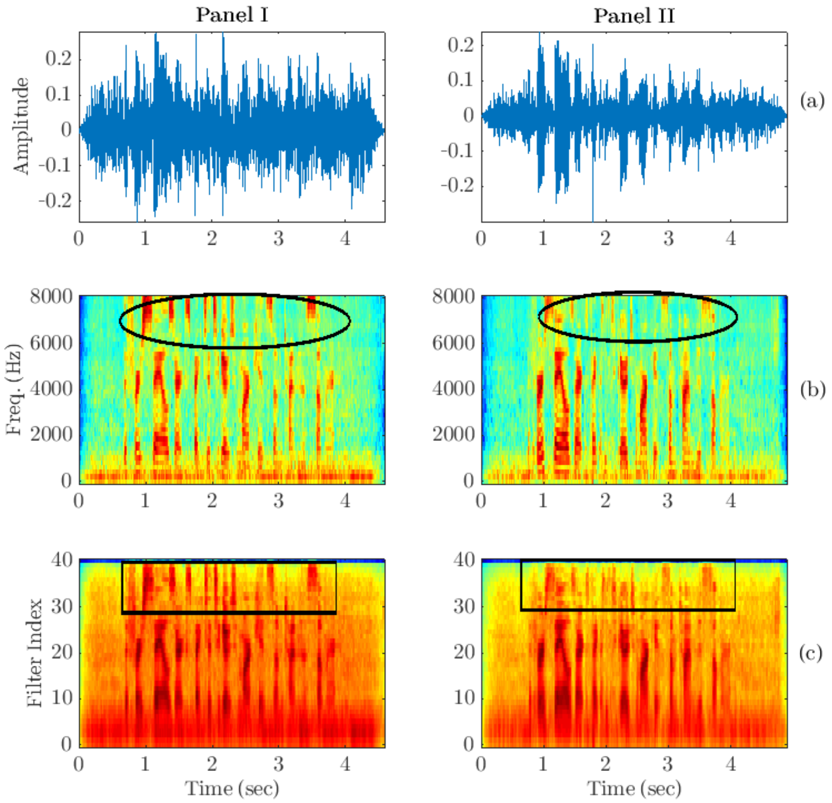


Figure 7.2: (a) Time-domain speech signal for (Panel I) real, and (Panel II) simulated, spectral energy density obtained from (b) Mel filterbank, and (c) GTFB. Highlighted ovals and box shows the spectral energy differences between (b) and (c). After [19].

7.2.1 Noise Suppression Capability of TEO

Historically, noise suppression capability of TEO was originally analyzed for near-field speech in car noise environment in [208] followed by our recent work on Wall Street Journal (WSJ) corpus [222]. Consider a clean speech signal, $s(n)$, degraded by a additive noise, $\eta(n)$, and the resulting signal is given as $y(n)$:

$$y(n) = s(n) + \eta(n). \tag{7.2}$$

The TEO of the noisy speech signal is given by:

$$\psi[y(n)] = \psi[s(n)] + \psi[\eta(n)] + 2\tilde{\psi}[s(n)\eta(n)], \tag{7.3}$$

where $\tilde{\psi}[s(n)\eta(n)]$ is the cross-ψ energy of $s(n)$ and $\eta(n)$. If $s(n)$ and $\eta(n)$ can be assumed statistically-independent, the expected value of $\tilde{\psi}[s(n)\eta(n)]$ is zero and hence, [164, 208],

$$E\{\psi[x(n)]\} = E\{\psi[s(n)]\} + E\{\psi[\eta(n)]\}. \tag{7.4}$$

We analyzed the Power Spectral Density (PSD) of a speech segment for far-field data (speech signals are taken from CHiME-3 corpus).

The PSD plots obtained from the (a) street, (b) pedestrian, (c) bus, and (d) cafe background environment obtained from with and without TEO is shown in Fig. 7.3 (from the CHiME3 corpus). Noise suppression capability of TEO which can be clearly observed in Fig. 7.3 [207, 222].

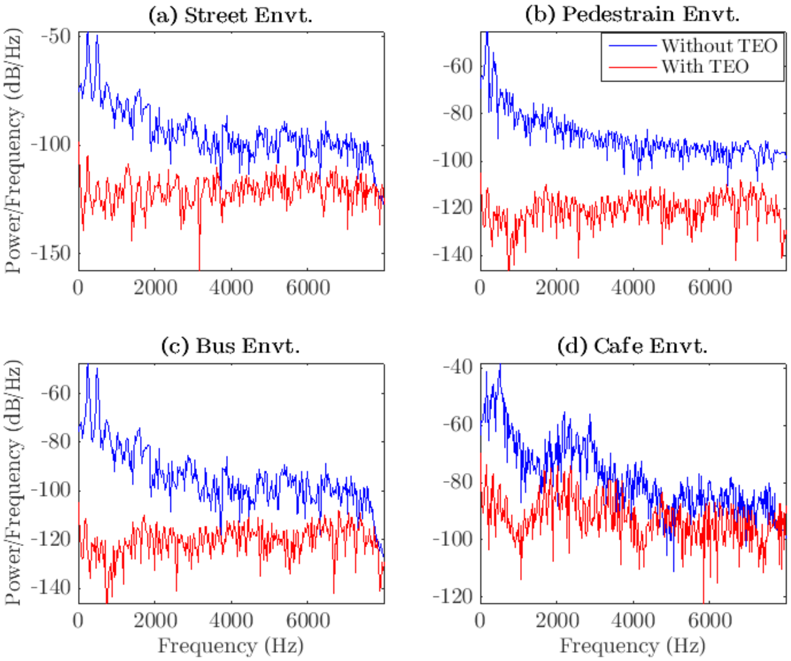


Figure 7.3: Power Spectral Density (PSD) of a Real Speech Segment of 20 ms with Street, Pedestrian, Bus, and Cafe Background from CHiME3 database. The PSD is Shown for Speech Segment With and Without TEO.

7.2.2 Experimental Setup

7.2.2.1 Near-Field and Far-Field ASR Corpus

The ASR experiments were performed on LibriSpeech and CHiME3 Corpus. The LibriSpeech task comprises English read speech data based on the LibriVox project [223]. The LibriSpeech database consists of two sets of clean speech data (100 hours + 360 hours), and noisy speech data (500 hours) for training. We used 100 hours of clean speech data to train the initial ASR model, and tested the trained models on test-clean and test-other. The statistics of the database is reported in [223]. In addition, we also performed experiments on CHiME-3 corpus which uses multi-microphone tablet device in everyday environments [221]. Four varied environments have been selected: Cafe (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). The real speech data is of *6*-channel recordings of the sentences from the WSJ0 (Wall Street Journal) corpus. The simulated data was developed by adding the clean speech utterances with the different environment in the background during recordings. For ASR evaluation, the corpus is divided into three subsets, namely, training, development, and test sets, respectively.

Table 7.1: Statistics of CHiME3 Corpus. After [19]

Corpus	**Real**	**Simulated**	**Environment**	**Speaker**	**Total**
Subset	(R)	(S)			**Utterances**
Train	1600	7138	-	4R-83S	8738
Dev	410	410	4	4	3280
Eval	330	330	4	4	2640

7.2.2.2 Feature Representation

For Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) training, MFCC features are extracted from the speech signals using a window length of 25 ms and shift of 10 ms. Delta and double-delta features are also appended resulting in *39*-dimensional (D) features. Human auditory system depends upon several thousands of subband filters, which results in dense filterbank in frequency-domain [207]. Hence, we performed experiments to investigate the significance of subband filters on ASR performance. The GTFB features are extracted with 60 subband filters using the process shown in Fig. 7.1.

The Fig. 7.4 shows the effect of increasing number of subband filtered signals during feature extraction. The experiments were performed with GMM-HMM,

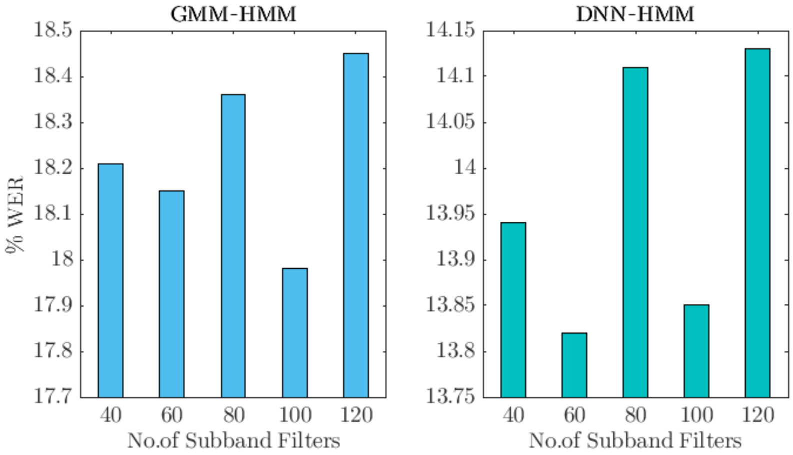


Figure 7.4: Effect of Subband Filtered Signals on GTFB Feature Set. After [19].

and DNN-HMM systems varying the number of subband filters from 40 to 120. It can be observed from the Fig. 7.4 that the features extracted using 60 number of subband filters are found to be relatively optimal on DNN-HMM systems compared to the other number of subband filtered signals and hence, further set of experiments for GTFB features were performed with 60 number of subband filtered signals. The KALDI toolkit is used to build the ASR systems for both the corpora (for details, please see the Appendix D) [224].

7.2.3 Experimental Results

7.2.3.1 Results for Near-Field ASR

To generate the alignments for training the DNN-based model, GMM-HMM system is used along with a tri-gram language model (LM) for decoding the ASR system. The experiments performed consists of *6* hidden layers in DNN with *1024* neurons (sigmoid activation) in each hidden layer. The output units are *3480* for each DNN, and the input is *11* frames (*5* left context, *1* current, and *5* right context) of *60*-D features concatenated together. The performance of the ASR system is analyzed using Word Error Rate (WER). The MFCC baseline using the LDA-MLLT (Linear Discriminant Analysis-Maximum Likelihood Linear Transform) system obtained WER of 12.28 %, and 34.92 % on test-clean and test-other dataset, respectively. The results obtained with GTFB are comparable to the MFCC feature set resulting in 12.71 % and 35.58 % WER on test-clean and test-other, respectively. The DNN-HMM system is trained on clean speech (using standard KALDI recipe) and tested in both clean and other conditions for MFCC and GTFB feature sets. The experimental results of test set using the DNN-HMM systems are reported in

Table 7.2.

Table 7.2: WER (%) on DNN-HMM System Trained on 100 Hrs of Training Data of LibriSpeech Corpus

Subset	**# Hrs**	**MFCC**	**GTFB**
Test-Clean	5.4	9.55	**9.40**
Test-Other	5.1	27.62	**27.54**

7.2.3.2 Results for Far-Field ASR

The baseline acoustic features are MFCCs (*13*-D). Three frames of left and right context are concatenated to form a *91*-D feature vector, which is compressed to *40*-D using LDA whose class is one of 2500 tied triphone HMM states. The tied states are modeled by a total of 15,000 Gaussians, MLLT, and feature space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT). The training set used is clean speech taken from WSJ0 corpus, and multi-noisy data and tested on noisy speech signals. On the other hand, the enhanced speech signals were tested on the clean speech and multi-enhanced speech signal. The DNN has 7 layers with 2048 units per hidden layer. The input layer has 5 frames of left, and right context (i.e., 11×40 = 440 units). The DNN is trained using the standard procedure: pre-training using restricted Boltzmann machines (RBM), cross-entropy training, and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion. In addition, the *N*-gram rescoring, and RNN-based LM is used for far-field ASR task. The experimental results with GMM, DNN, and RNN-LM-based ASR system are shown in Table 7.3, which is trained on multi-enhanced speech signal with MFCC and GTFB feature sets. The detailed results on different noises in CHiME-3 are reported in Table 7.4 with GTFB feature set. For all the noise conditions of CHiME-3 corpus, the GTFB feature set shows improvements over the baseline system especially on the evaluation set.

Table 7.3: WER (%) Using Beamforming and Enhanced Methods with Proposed Feature Set Trained on Multi-Enhanced Speech. After [19]

Method	Dev Set				Eval Set					
	Real		Sim		Real		Sim		Avg.(Real+Sim)	
	MFCC	GTFB	MFCC	GTFB	MFCC	GTFB	MFCC	GTFB	MFCC	GTFB
GMM-HMM	16.59	16.96	18.93	19.35	26.55	26.25	26.73	25.71	26.64	**25.98**
+ DNN (CE)	13.19	12.95	14.66	14.69	20.76	20.30	20.75	19.79	22.25	**20.04**
+ DNN (sMBR)	11.73	12.06	13.26	13.53	18.63	18.53	18.82	18.32	18.72	**18.42**
+ 5-gram rescoring	10.59	10.81	11.85	12.01	17.01	16.68	16.95	16.60	16.98	**16.64**
+ RNNLM	9.86	9.89	11.18	11.41	15.97	15.61	15.67	15.47	15.82	**15.54**

Table 7.4: WER(%) for Each Noise with GTFB Feature Set. After [19]

Envt.	Dev Set		Eval Set	
	Real	Sim	Real	Sim
Avg.	9.89	11.41	15.61	15.47
BUS	11.71	10.41	21.63	12.10
CAF	9.29	13.75	13.37	16.01
PED	8.16	9.59	13.66	16.16
STR	10.41	11.89	13.77	17.63

Table 7.5: Speech Recognition Performance (in % WER) on Near-Field and Far-Field Corpus. After [19]

Corpus	**System**	**Subset**	**SC**	**RI (%)**
LibriSpeech	Near-field	Test-Clean	**9.15**	**4.19**
LibriSpeech	Near-field	Test-Other	**26.24**	**4.99**
CHiME-3	Far-field	-	**14.68**	**7.20**

(SC: System Combination, RI: Relative Improvement)

Furthermore, in order to combine the possible complementary information from the amplitude and frequency of the speech signal, the posterior lattices obtained from the MFCC and GTFB feature sets were combined using lattice-level system combination (as shown in Table 7.5) [225]. The performance of the combined system is 9.15% and 26.24% for test-clean and test-other on LibriSpeech corpus, that results in relative improvement of 4.19% and 4.99% which is better than the MFCC alone. On the other hand, for CHiME-3 corpus, an relative improvement of 7.20% is obtained resulting in 14.68% WER. The individual noise condition performance after system combination (SC) of MFCC and GTFB on CHiME-3 corpus is reported in Table 7.6. Finally, we compared the performance of GTFB feature set with our proposed and other systems as reported in Table 7.7.

Table 7.6: WER (%) for Each Noise Condition With the System-Level Combination (SC) of MFCC and GTFB Feature Set. After [19]

Envt.	Dev Set		Eval Set	
	Real	Sim	Real	Sim
Avg.	9.43	10.79	14.78	14.59
BUS	11.37	9.96	20.56	11.11
CAF	8.70	12.58	12.29	15.11
PED	7.85	9.14	12.76	15.28
STR	9.79	11.46	13.50	16.87

In this Section, we explore the use of Teager energy spectral features-based

Table 7.7: Comparison With Other Systems on CHiME-3 Corpus (% in WER). After [19]

System	Dev Set		Eval Set	
	Real	Sim	Real	Sim
MFB [226]	11.6	14.3	22.6	25.5
PFB [226]	12.0	13.7	23.0	25.1
RAS [226]	11.8	14.6	21.6	23.1
MHE [226]	12.0	14.4	22.9	26.4
CVAE [226]	10.2	12.4	18.9	19.9
Ratemap+F_0 [227]	**5.51**	4.82	18.56	20.03
PNCC [228]	14.23	11.85	22.12	14.88
MESSL [229]	9.00	11.5	16.3	21.00
log-Mel [230]	12.58	10.66	23.86	20.17
DOC [230]	12.00	10.18	20.35	18.53
NIN-CNN [231]	10.64	11.21	**12.81**	18.47
DS Beamforming [232]	13.92	13.62	26.30	21.14
MFCC+RNNLM	9.86	11.18	15.97	15.67
GTFB	12.06	13.53	18.53	18.32
GTFB+RNNLM	9.89	11.41	15.61	15.47
GTFB+RNNLM_SC	9.43	10.79	**14.78**	**14.59**

acoustic model for near and far-field ASR tasks, where the GTFB feature set was extracted from Mel-spaced Gabor filterbank. The TEO preserves the amplitude and frequency modulation of a resonant signal, and it improves the time-frequency resolution. The noise suppression capability of TEO indeed helps for robust ASR task. The performance of the ASR system degrades when far-field speech is considered instead of near-field speech and hence, far-field system are more challenging, in particular, to handle background noise, reverberation. The experiments are performed on both LibriSpeech (near-field) and CHiME-3 (far-field) corpus. Significant reduction in % WER was achieved using the system combination using Minimum Bayes Risk (MBR) decoding of MFCC, and GTFB-based features.

7.3 Replay Attacks on Voice Assistants (VAs)

Voice might be the primary source of interface between humans and machines in the near future [32]. According to major companies that are involved in the speech recognition research believe that the perfect user interface, does not exist till date and to build it, knowledge of both sociology and technology fields are required [35]. The developed systems allows, one to wirelessly control lights, fans, TV, AC, security [36]. Home automation now-a-days is one of the major growing industries that changes the lifestyle of people [35]. Few of them target to have luxury and sophisticated platforms, on the other hand for special need, such as the person who are elderly and the disabled [37]. It is also useful for the people who live alone might require helping hand at home [37]. The Voice Assistants (VAs) considers that a cooperative speaker can be asked to pronounce a pre-defined sentence or phrase during both enrollment and test phases. This process is called text-dependent speaker verification as opposed to text-independent speaker verification in which no constraint is put on the input lexicon. In other words, text-dependent speaker verification can be defined as a speaker verification task in which the lexicon used during the test phase is a subset of the lexicon pronounced by the speaker during the enrollment.

Though the Voice Assistants (VAs) are convenience and ease, it raise new security issue because of their vulnerability to several types of spoofing attacks, such as replay, hidden voice commands, audio adversarial [8]. These attack pose a major threat, as they are easily conducted and hidden. These attacks can cause security access to several systems simultaneously. To protect the VAs approaches were proposed in [32, 233] by using the identification of the source speech samples, and rejecting the speech signals which comes from the machines. The work presented in [32, 233] is an extension of the anti-spoofing technologies for automatic speaker verification (ASV) system. The VAs usually use the far-field speech recognition, in addition, the acoustic environmental conditions are varied from indoor to outdoor. The feature vector also gets affected because of increase in distance, the noise-level increases rapidly and hence, it can be used for Spoof Speech Detection (SSD) task.

In this context, the organizers of the ReMASC (Realistic Replay Attack Microphone Array Speech Corpus) released publicly available database in order to focus on future research on the protection of VAs, and to make the *simulation-to-reality gap* lesser [8]. In this Chapter, we use Teager Energy Operator (TEO)-based feature for replay SSD task for VAs. The idea behind using the TEO is the nonlinear modeling of speech production. The TEO estimate the true total energy source of a resonance signal, and preserves both the amplitude and frequency information [52]. The supplementary information improves the time and frequency resolution and in addition, TEO has noise suppression capability that further helps to detect replay signal from its natural counterpart [51, 208, 222].

7.3.1 Feature Extraction

The block diagram of smooth Teager Energy Cepstral Coefficients (TECC) feature set is shown in Fig. 7.5. The TECC feature set is computed as per our earlier stud-

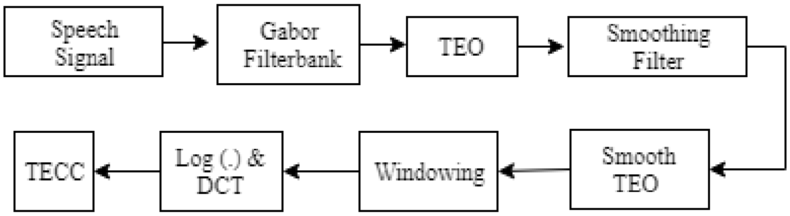


Figure 7.5: Block Diagram of Smooth Teager Energy Features. After [20].

ies [10, 11]. As the speech signal is the summation of several monocomponent signals, and Teager energy works on narrowband speech signal, we use Gabor filterbank to obtain subband filtered signals. The Gabor filters are asymmetric, non-constant-Q type, smoother, and broader compared to the Mels subband filters [55]. We use linearly-spaced Gabor filterbank to have almost equal bandwidth across all the frequency regions [13, 27]. It is also superior to the Mel filterbank, and provides higher robustness in situation when signal is degraded with additive noise and the other harmful interference (as shown in Fig. 7.7). Furthermore, these subband filtered signals are passed through the TEO to estimate the instantaneous Teager energy profile. These TEO profiles, in some segments observe energy spikes, and attributes to complex peaks. This makes to degrade the accurate detection of the signal and hence, to resolve the situation, the complex peaks and spikes are converted into smooth energy envelopes [234]. Here, Moving Average (MA) filter is used to smooth out complex energy peaks, and is given as [235]:

$$MA = filter(h, j, \Psi_d\{x_i[n]\}), \tag{7.5}$$

where h is the rectangle window with length L, and j is the constant that equals to 1. Furthermore, the smoothed Teager energy profiles are passed through the frame-blocking, and averaging using a short window length of 25 ms with a shift of 10 ms followed by logarithm operation to compress the data. Finally, Discrete Cosine Transform (DCT) is applied that has energy compaction property, and retaining first few DCT coefficients appended along with their delta and double-delta feature vector resulting in higher-dimensional smooth Teager Energy Cepstral Coefficients (TECC) feature set.

7.3.1.1 Effect of Smoothing Filter

Fig. 7.6(a) shows the segment of a speech signal, and its corresponding Teager energy profiles for natural (Panel I) and replay (Panel II) signals. The subband filtered Teager energy profile obtained without applying smoothing filter is shown in Fig. 7.6(b). Fig. 7.6(c) shows the smooth envelope of Teager energy profile,

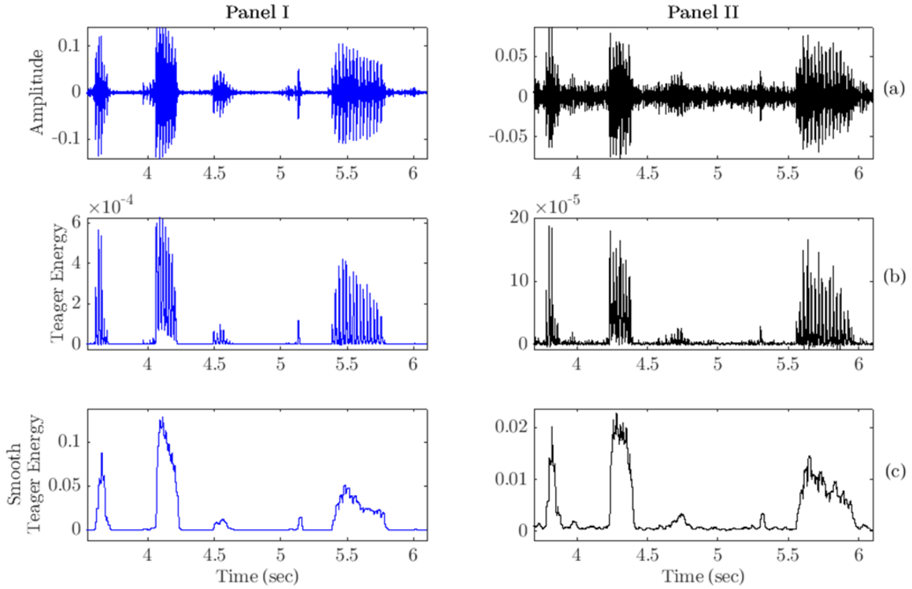


Figure 7.6: The Effect of Smoothing Filter on TEO. (a) Time-Domain Speech Segment Panel I: Natural and Panel II: Replay, (b) Corresponding Teager Energy Profiles, and (c) Teager Energy Profiles Obtained after Applying Smoothing Filter. After [20].

which is obtained after applying smoothing filter. It can be observed that after applying smoothing filter the Teager energy profiles do not carry spikes and complex peaks which is observed earlier in Fig. 7.6(b). In addition, the replay signal shows the energy fluctuations at the high peaks and small variation in the silence or unvoiced region. This observation is not present for the natural speech segment. Hence, the smoothing filter for TEO helps to detect the replay signal as compared to the TEO estimated from without applying smoothing filter (refer Table 7.11).

7.3.1.2 Spectral Analysis

The time-domain speech and corresponding spectral energy densities obtained from Mel and Teager energy filterbank are shown in Fig. 7.7 for natural (Panel I) and replay (Panel II) speech signals. It can be observed that for the natural speech signal the spectral energy do not vary much apart from its intensity for Teager energy spectral features. For the replay speech signal, we can see that the formant patterns obtained from the Mel spectral energy are distorted, and fails to capture the energy content of the signals. However, the spectral features obtained from the Teager energy preserve this formants and harmonics characteristics of the replay signals. This is an initial observation and difference between the natural and replay signal obtained on the ReMASC database.

The PSD of the natural and replay signal from the ReMASC database is as

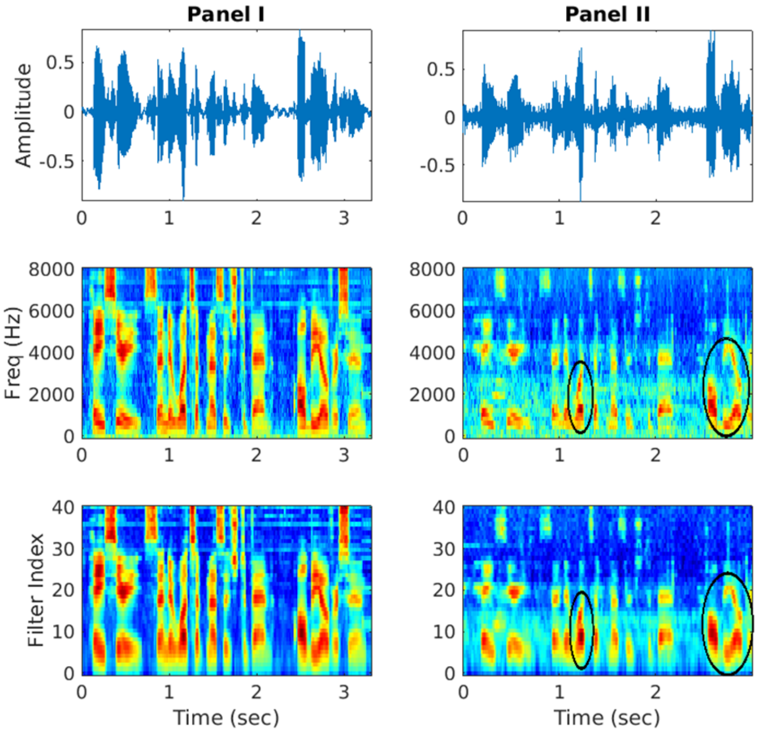


Figure 7.7: (Top) Time-Domain Speech Signal for Natural and Replay (Panel I and II), Corresponding Spectral Energy Density for Mel Filterbank and Teager Energy (Middle and Last Row). After [20].

shown in Fig. 7.8. PSD is obtained for a speech segment, where the segment is obtained by applying TEO (black line) and without applying TEO (blue line). TEO has good noise suppression capability which is observed as shown in Fig. 7.8. In particular, for natural speech segment Fig. 7.8 (left side) the difference is visible in lower as well as in higher frequency regions. Similarly, in case of replay speech segment as shown in Fig. 7.8 (right side), the difference goes on decreasing as the frequency increases from lower-to-higher frequency. In particular, the difference present in lower frequency shows huge gap that helps to detect replay signal from its natural counterparts.

7.3.2 Experimental Setup

The ReMASC database is classified into two sub-sets, namely, core set (30k number of samples), the and quick evaluation set (2k number of samples covering all the recording conditions) [8]. The experiments performed in this Section are done on quick evaluation set. The statistics of the ReMASC database is given in [8]. Following systems were compared with smooth TECC feature set for the classification of natural *vs.* replay speech signal.

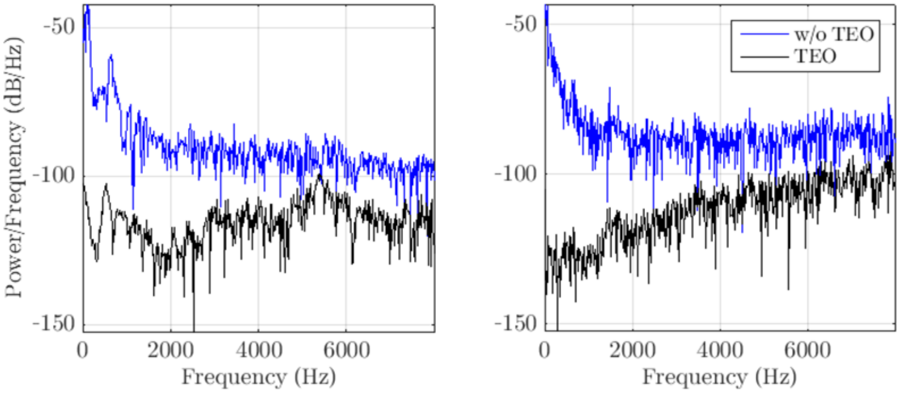


Figure 7.8: The PSD of the Natural (Left) and Replay (Right) Speech Segment Obtained by Applying (Black Line) and Without Applying (Blue Line) TEO. After [20].

Table 7.8: The Statistics of ReMASC Database (* Indicates Incomplete Data Due to Recording Device Crashes). After [20]

Environment	# Subjects	# Genuine	# Replayed
Outdoor	12	960	6900
Indoor 1	23	2760*	23104
Indoor 2	10	1600	7824
Vehicle	10	3920	7644
Total	55	9240	45472

Baseline Systems: The Constant Q Cepstral Coefficients (CQCC) features were extracted with 30 static coefficients (with log-energy), resulting in total *90*-dimensional (D) feature vector (including *30*-Δ and *30*-ΔΔ).

MFCC: The MFCC feature set were extracted using 40 Mel filterbank with f_{min}=10 Hz, and f_{max}=8000 Hz. We obtain *13*-D static features appended along with their Δ and ΔΔ coefficients resulting in *39*-D feature vector.

TECC: The TECC feature set were extracted using 40 linearly-spaced Gabor filterbank with f_{min}=10 Hz, and f_{max}=8000 Hz. For each subband filtered signals, we obtain *40*-D static features appended along with their Δ and ΔΔ coefficients resulting in *120*-D feature vector. The smooth TECC (TECC_sm) features were extracted similar to TECC features.

Gaussian Mixture Model (GMM) is more popular and well known classification technique widely used in signal processing and pattern recognition literature. GMM is a generative model which represent each class as a weighted sum of M multivariate Gaussians, and is given by $p(x|\lambda) = \sum_{k=1}^{M} w_k p_k(x)$, where w_k is the

k^{th} mixture weight, and $p_k(x)$ is a D-variate Gaussian density function with mean vector μ_i and covariance matrix Σ_i. The model parameter is defined by λ. The decision of whether the test speech being natural or spoofed depends on the scores of Log-Likelihood Ratio (LLR) $= log\frac{P(X|H_0)}{P(X|H_1)}$, where $P(X|H_0)$ and $P(X|H_1)$ are the likelihood scores of natural and replay speech, respectively. The score-level fusion is performed to combine possible complementary information, and is given by $LLK_{fused} = \alpha LLK_{feature1} + (1-\alpha)LLK_{feature2}$, where $LLK_{feature1}$ and $LLK_{feature2}$ is log-likelihood score of feature1 and feature2, respectively. The fusion parameter (α) lies between $0 < \alpha < 1$ is tuned as weight factor.

7.3.3 Experimental Results

The experiments were performed on ASVspoof 2017 V2.0 challenge [5] and ReMASC database [8]. For ReMASC database, the experiments are divided into three tasks, i.e., training done on: RedDots Pretrained database, Environment-Independent, and Environment-Dependent. The details of three different task are discussed below:

7.3.3.1 Task 1: RedDots Pretrained

In this experimental Section, we train the model using RedDots Replayed (ASVspoof 2017 version 2 challenge) dataset (i.e., training + development set), and tested on the quick evaluation set of ReMASC dataset. This experiment is to evaluate the performance of mismatch training and testing condition, as the training was done using the data recorded in different acoustic environments. The results using different feature sets are shown in Table 7.9 for all the four acoustic environments. It can be observed that the performance degrades resulting in high Equal Error Rate (EER) around 50 % for all the environments, which is not a good case for ASV system. This high error is due to mismatch in training and testing database.

Table 7.9: % EER on RedDots Pretrained Conditions. After [20]

Feature Set	EER			
	Env_A	Env_B	Env_C	Env_D
CQCC	49.41	45.36	38.11	45.81
MFCC	47.65	**28.95**	42.36	**33.87**
TECC	47.05	37.67	**36.64**	43.28
TECC_sm	**46.09**	40.03	38.36	43.13

TECC_sm indicates TECC smooth

7.3.3.2 Task 2: Environment-Independent

In this Section, the experiments are performed on environment-independent condition on ReMASC database. In particular, the performance is evaluated, when the training models were developed on a particular set of environments and tested on unseen target environment. Specifically, the system models were developed using data of three environments and tested on the target environment. The results for different feature sets are shown in Table 7.10 for all the four acoustic environments. It can be observed that the performance is better than the RedDots Pretrained model (except Env-D). However, the results are still not satisfactory, in particular, for complex noise (Env-C & D), and when the signals are recorded at different positions (Env-B). This concludes that the acoustic environment, and recording do have high impact during training and testing of data samples.

Table 7.10: EERs on Environment-Independent Conditions. After [20]

Feature Set	EER			
	Env_A	Env_B	Env_C	Env_D
CQCC	39.70	**33.82**	**33.16**	49.64
MFCC	**39.55**	37.22	45.60	50.00
TECC	47.59	34.43	36.48	50.00
TECC_sm	48.99	37.54	41.48	**47.81**

7.3.3.3 Task 3: Environment-Dependent

In this set of experiments, the performance evaluation was done on the environment-dependent data. Specifically, the training models were generated using all the acoustic environment from the core dataset and tested on the quick evaluation set of ReMASC database. The training dataset were randomly selected and were speaker-independent (i.e., speakers selected for training data were not present during testing). We observe improvement compared with the RedDots Pre-trained and the environment-independent model. This indicates that the information of data present during training strengthens the performance even for unknown speaker. The results with all the feature sets for environment-dependent task are shown in Table 7.11. In addition, we also performed score-level fusion of CQCC and MFCC features with smooth TECC feature set, that further improved the performance of SSD task compared to its individual performance.

Table 7.12 shows the average EER over all the environments and compared smooth TECC feature sets results with CQCC and MFCC feature set. It can be observed from the Table 7.12 the % EER obtained with smooth TECC feature set

Table 7.11: % EERs on Environment-Dependent Conditions. After [20]

Feature Set	% EER			
	Env_A	Env_B	Env_C	Env_D
A:CQCC	19.7	10.31	9.94	9.62
B: MFCC	30.74	5.63	27.02	8.81
TECC	20.27	6.85	11.24	9.43
C: TECC_sm	18.76	8.65	10.43	10.09
A+C	**16.57**	7.84	**8.10**	8.60
B+C	18.71	**5.63**	10.43	**8.17**

performed better (11.98 % EER) compared to the other feature sets. The score-level fusion of CQCC and smooth TECC further reduce the EER to 10.27 %. Fig. 7.9 shows the EER of CQCC, MFCC, and TECC feature set on ASVspoof 2017 challenge version 2.0 and ReMASC database. It can be observed from the plot that EERs have large gap for CQCC and MFCC feature set as the database is changed. With TECC feature set, less difference in EER is obtained and hence, it is more generalized feature set compared to the other approaches.

Table 7.12: Overall Average EER for All Tasks and Tested on Quick Evaluation Dataset. After [20]

Condition	Average EER(%)				
	CQCC	MFCC	TECC_sm	A+C	B+C
Task 1	44.67	38.20	41.16	-	-
Task 2	39.08	43.09	43.95	-	-
Task 3	12.41	18.05	11.98	**10.27**	10.73

In this Section, we studied the importance of smooth Teager energy profiles for replay SSD task. In particular, the small amplitude variations for replay signals recorded in different acoustic environments helps to discriminate between natural and replay signals, which is not present in natural signal. The Teager energy have the noise suppression capability and this difference is clearly visible in replay signal because the replay signal carries noise along with clean speech. The performance degrades when the training and testing samples are not from the same acoustic background. In addition, it also fails when the testing is done for environment-independent condition. The smooth TECC feature set performed better compared to the baseline and MFCC features resulting in 11.98 % EER. Furthermore, reduction in % EER was achieved using the score-level fusion of baseline and smooth TECC resulting in 17.24 % relative improvement over the baseline system.

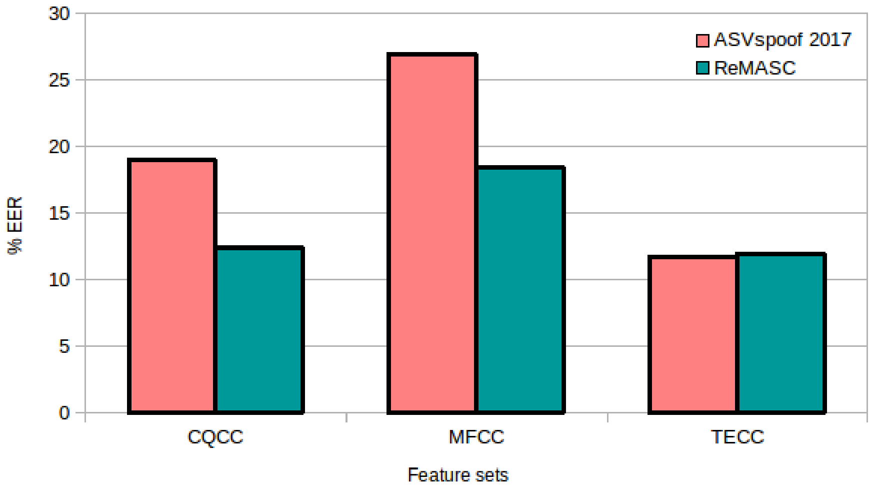


Figure 7.9: Comparison of CQCC, MFCC, and Smooth TECC on ASVspoof 2017 Challenge and ReMASC Database. After [20].

7.3.4 Energy Separation Algorithm (ESA)-Based Features

In addition to the above proposed feature set, we also performed the experiments using Energy Separation Algorithm-Instantaneous Amplitude Cepstral Coefficients (ESA-IACC), and Energy Separation Algorithm-Instantaneous Frequency Cepstral Coefficients (ESA-IFCC) feature extraction process is shown in Fig. 7.10.

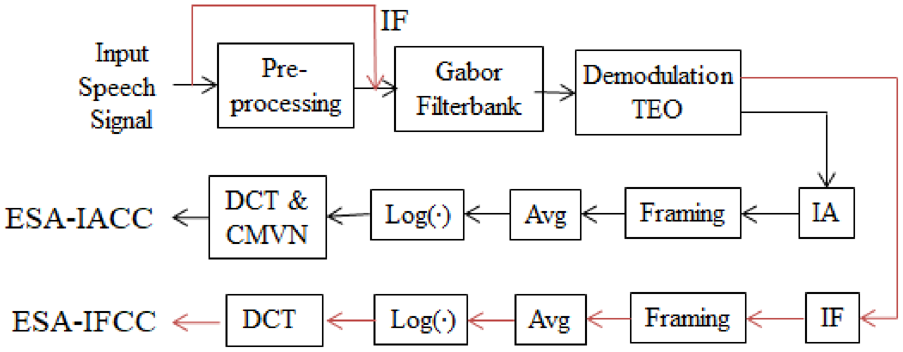


Figure 7.10: Block Diagram of ESA-IACC, and ESA-IFCC Feature Sets. After [21, 27].

The input speech signal is passed through the pre-emphasis filter to enhance the higher frequency regions. Furthermore, the speech signal is passed through the linearly-spaced Gabor filterbank in order to obtain narrowband filtered signals in order to estimate the instantaneous Teager energy profile along with ESA [51, 166, 167, 170]. The ESA provides the corresponding Instantaneous Amplitude and Instantaneous Frequency (IA-IF) components of a narrowband filtered signals [190]. These estimated IA and IF components are further processed to obtain corresponding speech segments with a window length of 25 ms along with a shift of 10 ms followed by logarithm operation to compress the data. To obtain a low-dimensional representation that has compact energy, Discrete Cosine Transform (DCT) is applied along with Cepstral Mean Variance Normalization (CMVN) to reduce the channel mismatch/distortion conditions [174]. Finally, retained few DCT coefficients, i.e., ESA-IACC , and ESA-IFCC appended along with their Δ and $\Delta\Delta$ features to obtain higher-dimensional feature vector. Please note that the experiments performed in Section 7.3.5 for ESA-IFCC feature set are extracted without applying pre-and-post-processing, as this process gave better results than the other feature extraction parameters used.

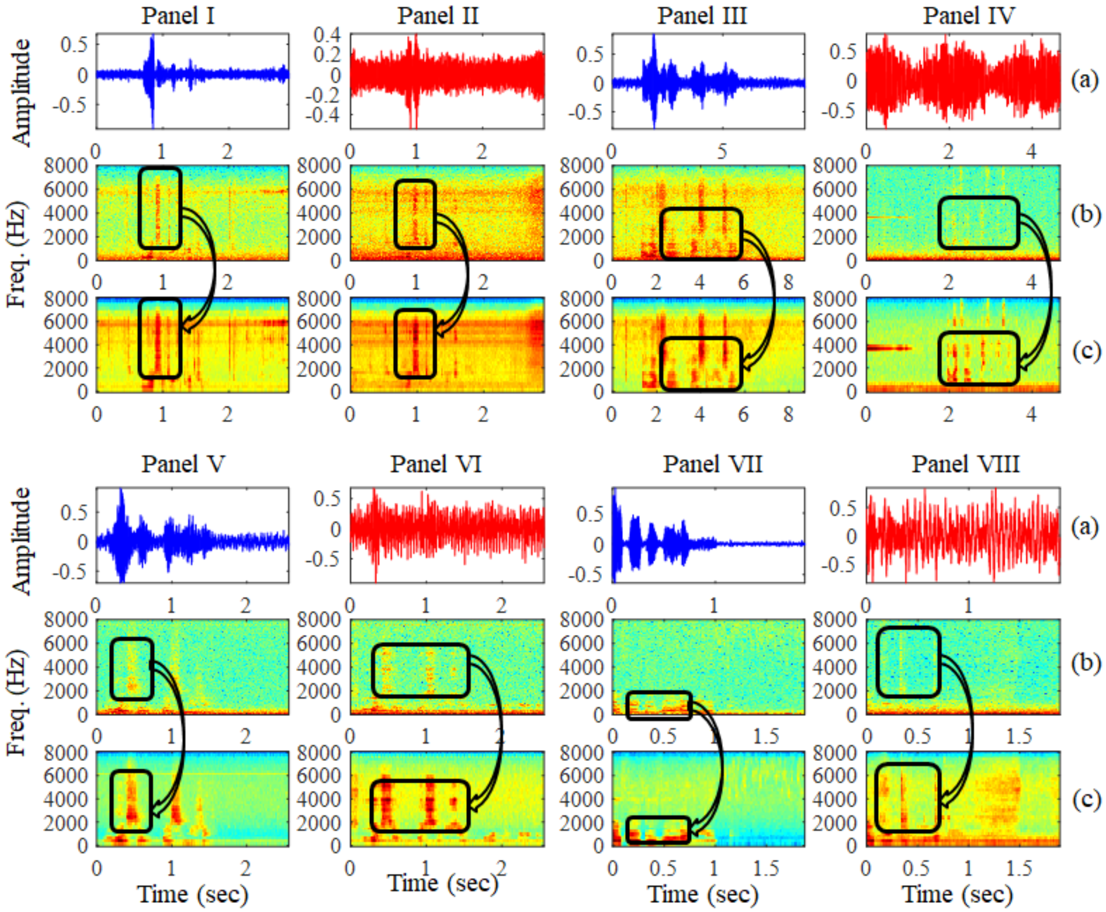


Figure 7.11: Spectral Energy Densities Obtained from the Traditional STFT, and Teager Energy-Based Approach for Different Acoustic Environments. Panel I-II: Outdoor, Panel III-IV: Indoor 1, Panel V-VI: Indoor 2, Panel VII-VIII: Vehicle. (a) Time-Domain Signal in Different Acoustic Environment Along with Their Corresponding, (b) Traditional STFT, and (c) Teager Energy-Based Approach. After [21].

In addition, we also observed and compared the spectral energy densities of traditional Short-Time Fourier Transform (STFT) spectrogram with the spectral energy obtained from the Teager energy-based approach as shown in Fig. 7.11.

The comparison is shown for all the acoustic environments from the ReMASC database, in particular, outdoor (Panel I-II), indoor 1 (Panel III-IV), indoor 2 (Panel V-VI), and vehicle (Panel VII-VIII). For outdoor environment, it can be observed from the Fig. 7.11 that the spectral energy is not preserved for both Panel I and Panel II. In addition, for high frequency regions, we observe high energy. However, with Teager energy-based approach, we preserve the mid and higher frequency information compared to the traditional spectrogram. The spoof signal of corresponding outdoor environment shows much more distortion in spectral energies compared to its natural counterpart. As the recording is done in the open outdoor area, it is indeed possible that the signal carries different types of noise along with it and hence, the performance degrades for outdoor environment. Similarly, we observe the spectral energy differences for other acoustic environment. In particular, for indoor 2 and vehicle acoustic environments, it can be clearly observed that the spectral energy obtained from the Teager energy-based approach preserves much more information about the formants and harmonics compared to the traditional spectrogram and hence, the performance for these environment is better compared to the other environments (discussed in Section 7.3.5).

7.3.5 Experimental Results

This Section describes the experiments performed on the Task 3, i.e., environment-dependent. We observed the Power Spectral Density (PSD) obtained after applying TEO on the small speech segment for the natural, and its corresponding spoof speech signal as shown in Fig. 7.12. The PSD for different environment, namely, (a) outdoor, (b) indoor 1, (c) indoor 2, and (d) vehicle shows the difference from its natural counterpart. In particular, we observe differences in the PSD plots for the indoor 2 and vehicle environments, which indeed help us to detect the spoof signal (which is also observed from our experimental results discussed in the next sub-Section). Furthermore, the performance for ESA-IACC, and ESA-IFCC feature sets is enhanced in the next sub-Section.

7.3.5.1 Results on ESA-IACC Feature Set

The experiments are performed by varying the number of subband filters in a Gabor filterbank from 40 to 100 for all the acoustic environments (as shown in Fig. 7.13). It can be observed from Fig. 7.13 that the effect of number of subband filters indeed degrades the performance for a particular environment, and at the same time, it performs better for the other environments. In particular, for Env

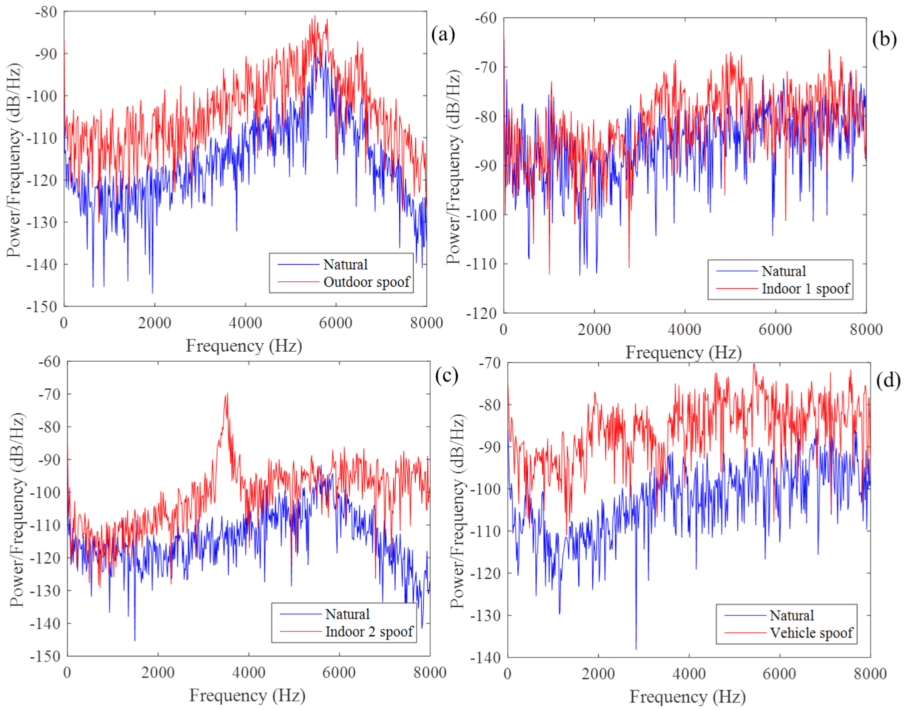


Figure 7.12: PSD of Natural (Blue Color) Speech Segment, and its Corresponding Replay (Red Color) Speech Recorded in (a) Outdoor, (b) Indoor 1, (c) Indoor 2, and (d) Vehicle Acoustic Environment. After [21].

B (indoor 1), it can be observed that we get high EER for all the number of subband filters. The possible reason behind it could be the environmental conditions, which the replay spoof speech is recorded. Since indoor 1 environment is quiet study room, the replay spoof signal will be similar to the natural speech resulting in less discrimination in replay speech from its natural counterpart, thereby degradation in the SSD performance. On the other hand, the spoof signal when recorded in other environments, i.e., indoor 2 and vehicle are able to detect as these environments are having noise added in the replay signals that is used as the discrimination feature from the natural speech because the Teager energy-based features have noise suppression capability, and thus, its features are robust to noise sensitivity.

7.3.5.2 Results on ESA-IFCC Feature Set

Similar to experiments in Section 7.3.5.1, we performed the experiments for ESA-IFCC feature set with varying the number of subband filters from 40 to 100. It can be observed from Fig. 7.14 that for acoustic environment indoor 2 and vehicle, the EERs are low compared to the other two acoustic environments, namely, outdoor and indoor 1. With increasing the number of subband filters in a filterbank, the EER decreases and hence, it proves that the narrowband filtering for extracting

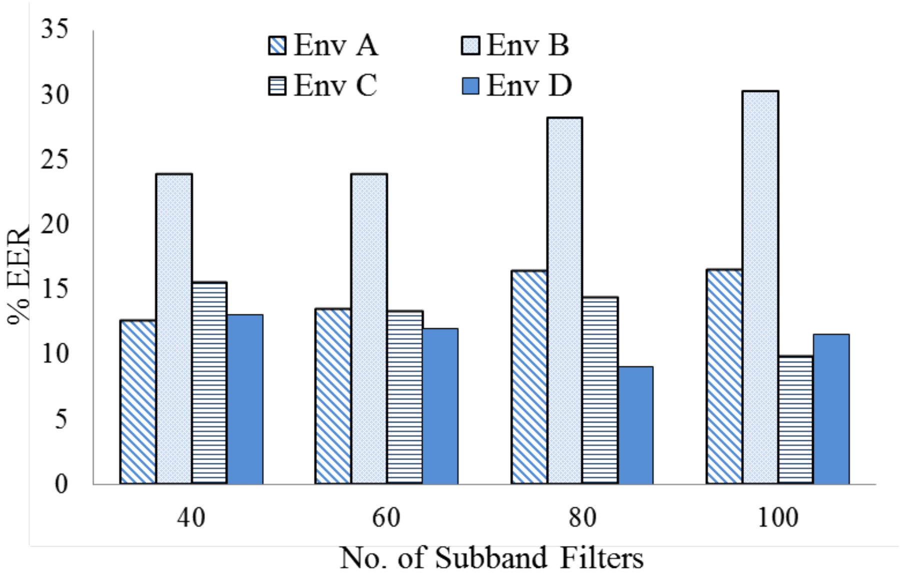


Figure 7.13: Results in EER (%) for ESA-IACC Features Sets with Varying the Number of Subband Filters for Different Acoustic Environments. After [21].

TEO-based features are essential for detecting spoof speech signals, which also further depends on the acoustic environment (i.e., noisy *vs.* clean environment).

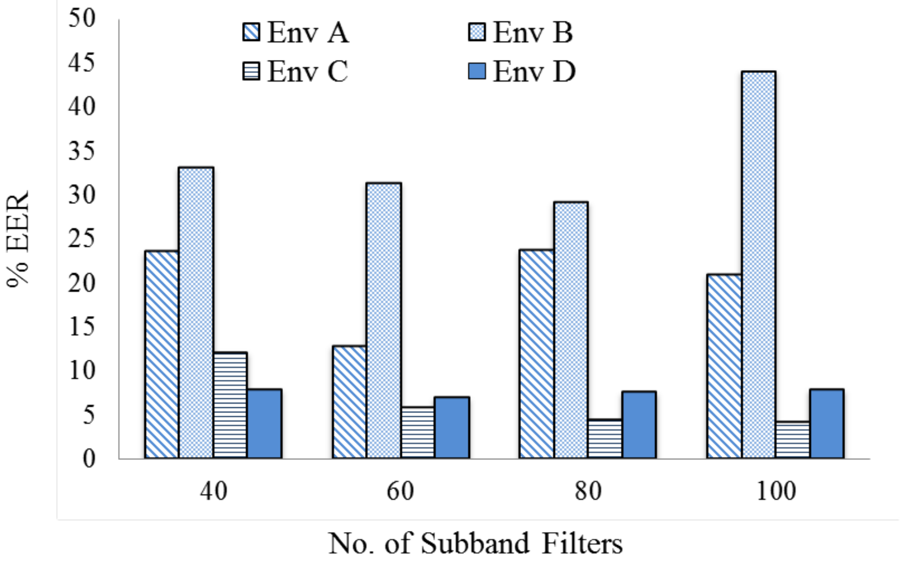


Figure 7.14: Results in EER (%) for ESA-IFCC features sets with varying the number of subband filters for different acoustic environments. After [21].

7.3.5.3 Results with Score-Level Fusion

We further compared results for our proposed feature sets with the baseline system along with LFCC feature sets in Table 7.13. It can be observed that the ESA-IACC and ESA-IFCC feature set performed better for Env A, Env C, and Env D. However, for Env B, the ESA-based feature sets fail to detect the replay speech signal. When compared to CQCC and LFCC feature sets, we obtained much lower

EERs for all the environments apart from indoor 1. Furthermore, we performed the score-level fusion of ESA-IACC and ESA-IFCC feature sets to further improve the performance of the replay SSD task. The score-level fusion indeed helped to get the lower EER than the individual EERs for both the feature sets. For outdoor, indoor 2, and vehicle environment, the score-level fusion gave EER of 11.92 %, 2.07 %, and 5.18 %. It represents that the score-level fusion of both the feature sets capture complementary information that helped us to improve the replay SSD performance than the individual feature sets alone.

Table 7.13: Comparison (in % EER) with Other Feature Sets Along with Score-Level Fusion Results (in % EER). After [21]

Feature sets	**% EER**			
	Env A	**Env B**	**Env C**	**Env D**
CQCC	15.26	**17.41**	6.15	6.59
LFCC	22.44	24.41	15.97	18.24
ESA-IFCC	19.36	29.11	4.06	6.22
ESA-IACC	12.59	23.84	9.81	9.11
ESA-IFCC+ESA-IACC	**11.92**	21.00	**2.07**	**5.18**

In this Section, we studied the importance of different acoustic environments for replay attack detection on Voice Assistants (VAs). In particular, we found that the noisy and clean environment indeed affect the performance to detect the replay speech signal from its natural counterparts. We used Energy Separation Algorithm (ESA)-based Instantaneous Amplitude and Instantaneous Frequency feature sets to detect the replay signals. The speech signal when recorded in noisy environment has distortions, however, using the ESA-IFCC feature sets; this type of replay signals are classified from it's natural counterpart. On the other hand, when the signals are recorded in the clean environment, they are difficult to detect as they might be similar to the natural signal and hence, very less differences in them are observed. Thus, for the clean environment, our proposed feature sets fails to classify the replay signal and hence, more detailed analysis and study is required to detect the replay signal in such scenarios, which forms our immediate future work.

7.4 Whisper Speech Detection (WSD)

Whisper speech detection has become a topic of research interest. The differences in whispered *vs.* normally-phonated speech are primarily due to the noisy structure, lower Signal-to-Noise-ratio (SNR), absence of glottal vibrations, shift in formant structures, etc. [236]. In Automatic Speaker Recognition (ASR) systems, the training and testing are performed on normal and corresponding whisper, i.e., mismatch speech dataset and hence, the performance degrades. Several approaches have been proposed to attenuate the mismatch through feature transformations [237], model adaptation [238–241], or using alternative sensing technologies, such as throat microphone [242].

Whispered Speech Recognition (WSR) is an active field of research which is obstructed by the lack of systematically, and suitable collected corpora. There are few publicly available databases for parallel normal and corresponding whispered speech collected for different languages, such as English [243, 244], Mandarin language [245], Japanese [238], and Serbian language. However, the vocabulary of these database are small or medium-sized and only few of them have the corresponding transcription and are phonetically-balanced. One of the first experiments on automatic whisper recognition is reported in [238]. The key goal was to develop a speech recognizer which is specifically capable of handling whisper on cell phones in noisy conditions. Using MFCC-HMM system, they analyzed different mismatched train/test scenarios by taking three speech modes, namely, whisper, low-voice speech, and neutral speech. Severe degradation in ASR was reflected due to use of mismatch data. However, there was outstanding result when ASR model is trained on whisper (whisper speech model), and also it was working well enough for testing with either type of speech. Next, it was also observed that covering the mouth and the cellphone with a hand can increase of SNR in noisy environment to a certain extent.

Unlike normal speech, whisper speech does not contain fundamental frequency (*Fo*) due to the absence of voice harmonic distortion, and formant shifting in the lower frequency regions [238, 246]. In [246, 247], it has been observed that normal and whisper speeches have different formant characteristics, where vowels of Serbian and English language were used. It was observed that the formant frequency F_1 for whisper speech is greater than that of normal speech for both female and male speakers. In addition, F_2, F_3, and F_4 shifts depends on the type of vowels, and they do not exhibit consistent trends [248]. This characteristics can be used as a main attribute to classify the normal *vs.* whisper speech. The same study explored that formant bandwidths for whisper vowels has a general expansion than that of vowels.

To improve the performance of ASR system, recently various approaches have been proposed for the conversion of an whispered speech to normal speech with the aim of improving speech intelligibility, and naturalness [249–254]. As the use of voice assistants (VAs), and Text-to-Speech (TTS) systems is becoming more common and hence, the need for the speaker to interact with such systems privately is also increasing simultaneously. In such realistic scenarios, a user may wish to whisper to the device, and would also expect a response in a whispered voice, as is the case with the recently released version of *Amazon Alexa* [255]. To further improve robustness of these ASR systems, some pre-processing can be performed by developing clusters of normal, and whisper speech so that they can be identified beforehand, and further processing can be done accordingly, in particular, ASR of whispered speech [256]. Due to this reason, the classification of these two types of speeches becomes an important part of ASR systems, especially in the context of commercial success of IPA or VAs.

7.4.1 Acoustic Features Used

7.4.1.1 LFCC *vs.* MFCC

While extracting MFCC or LFCC feature sets, the speech signal is windowed and DFT is computed for each frame to get the Short-Time Fourier Transform (STFT), $X(n, \omega_k)$. The energy in STFT is weighted by each Mel scale filter frequency response, $V_l(\omega)$, to get the l^{th} energy coefficient, i.e.,

$$E_{mel}(n,l) = \frac{1}{A_l} \sum_{k=L_l}^{U_l} |V_l(\omega_k) X(n,\omega_k)|^2. \tag{7.6}$$

The real cepstrum C_{mel} associated with the $E_{mel}(n,l)$ is referred to as MFCC:

$$C_{mel}[n,m] = \frac{1}{R} \sum_{l=0}^{R-1} log(E_{mel}(n,l))cos(\frac{2\pi}{R} lm), \tag{7.7}$$

where R is the number of subband filters. The transformation in eq. (7.7) is also known as Discrete Cosine Transform (DCT). In this Chapter, we considered MFCC as the baseline (state-of-the-art) feature set to compare the result [175, 257]. Both MFCC, and LFCC use similar algorithm for feature extraction except the type of frequency response used in order to obtain the weighted sum from the spectrum. In general, Mel scale gives more information (resolution) to the lower frequency regions, and less information to the higher frequency regions [258]. This arrangement suggests that the MFCC fails to extract effective spectral characteristics at the high frequency regions. Both MFCC and LFCC feature sets use triangular-shaped filters in order to obtain the subband filtered components. This means that the features which can retain both low frequency, and high frequency characteristics could be effective for classification of whisper *vs.* normal speech.

7.4.1.2 TECC

So far, nonlinear TEO-based features introduced very promising results in ASR of quiet and unvoiced murmured speech as well in speech classification under stress and noisy conditions [207, 259]. On the other side, the characteristics of whispered speech might be considered to have similarities with non-audible murmur, noise-corrupted speech and speech under stress. Due to these similarities, it was expected that the TECC could be good descriptors of whispered speech.

The basics of TECC feature extraction process is similar to the MFCC, however, it is having one major difference in terms of estimating the energy. For TECC feature set, the nonlinear TEO is employed that estimates the instantaneous Teager energy instead of standard energy (Squared Energy Operator (SEO) that employ L^2 norm of a signal) [260]. Other details of TEO and TECC feature sets are given in Chapter 3.

The nonlinear modeling mechanism of the speech production is the main motivation behind using TEO instead of standard SEO is [52]. In traditional linear acoustic theory, it is assumed that the airflow from the vocal tract system propagates as a plane wave. However, this assumption may not hold for real speech signal as it is produced due to vortex flow interactions, which are nonlinear in nature [261]. Since, the whispered speech is nonlinear and contains extreme turbulent airflow, TEO provides an efficient way for signal processing due to its inherent capability to capture properties of airflow pattern. The TEO incorporates both amplitude and frequency information and computes the 'true' total source energy of a resonance signal [207] along with improving time-frequency components of rapid energy changes [260]. To obtain a low-dimensional representation that has compact energy, Discrete Cosine Transform (DCT) is applied along with Cepstral Mean Normalization (CMN) (also known as Cepstral Mean Subtraction (CMS)) in order to reduce the channel mismatch/distortion conditions [174]. Finally, retained few DCT coefficients to get Teager Energy Cepstral Coefficients (TECC) which are appended along with their Δ and $\Delta\Delta$ features to obtain higher-dimensional feature vector (for more details, please refer Chapter 3) [11].

In addition, we compared the spectral energy densities (as shown in Fig. 7.15) obtained from the traditional Short-Time Fourier Transform (STFT), and Teager energy-based approach for the both corpora, namely, wTIMIT and CHAINS. In particular, Panel I and Panel II in Fig. 7.15 shows the natural and corresponding whisper speech for wTIMIT corpus and Panel III and Panel IV shows for

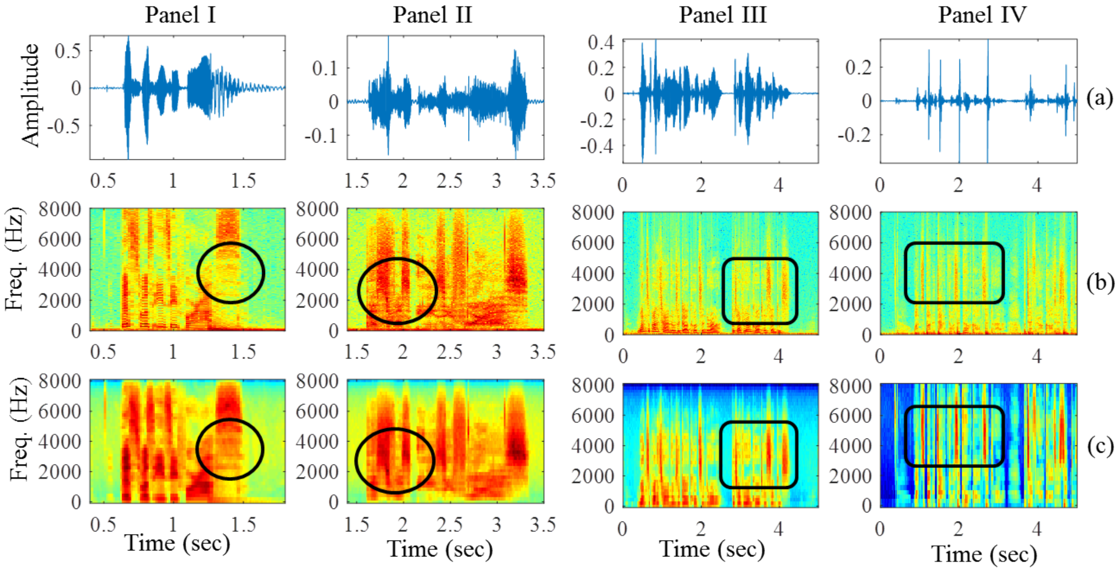


Figure 7.15: Panel I and Panel II are the Natural and Corresponding Whisper Speech from the wTIMIT Corpus, Panel III and Panel IV are the Natural and Corresponding Whisper Speech from the CHAINS Corpus. (a) Time-Domain Speech Signal, (b) Traditional STFT Spectrogram, and (c) Spectral Energy Density Obtained from Teager Energy-Based Approach. The Discriminative Regions are Indicated by Circle and Box for Corresponding wTIMIT and CHAINS Corpora. After [22].

natural and whisper speech for CHAINS corpus. The spectral energies for the time-domain signal for traditional STFT and Teager energy-based approach are shown in Fig. 7.15 (b) and Fig. 7.15 (c), respectively. It can be observed that the energy density obtained from the Teager energy-based approach preserves much more information in low as well as in high frequency regions as compared to the traditional spectrogram. In particular, the formants are well preserved for the natural speech (refer Panel I (b)) which are not visible for the traditional spectrogram for wTIMIT corpus. In case of CHAINS corpus, it is observed that many of the higher frequency information are not preserved when estimated from the traditional spectrogram whereas the Teager energy-based approach do carry the information in the higher frequency regions. This spectral energy obtained from the Teager energy-based method indeed help to classify the whisper speech from its natural counterpart.

7.4.2 Experimental Setup

7.4.2.1 Corpora Used

We performed the experiments on two corpora, namely, wTIMIT and CHAINS. The wTIMIT corpus was collected in two phases, the first phase was recorded in

Singapore, and the second phase was recorded in the USA [239]. In this chapter, we used only data recorded in USA. The sampling frequency of data is set as 44.1 *kHz*, and all the recordings were done in clean acoustic environment. For training, we have 9219 and 11325 for normal and whisper utterances whereas 412 whisper utterances, and 727 normal utterances are used for testing. The CHAINS corpus is designed to characterize speakers as individuals. The corpus contains the recordings of 36 speakers (20 male and 16 female) in two different sessions with a time separation of about two months. For training, we considered 1036 normal and whisper utterances, and 296 normal and whisper utterances for testing.

7.4.2.2 Feature Extraction Parameters

All the utterances during feature extraction process were first resampled to 16 *kHz* from 44.1 *kHz*. This is done primarily so as to reduce the number of samples thereby saving computational cost. The process of frame-blocking is carried out by taking a window length of 20 *ms* with an overlap of 10 *ms*. We have considered *39*-D feature vector extracted from 40 number of subband filters in a filterbank for MFCC, LFCC, and TECC feature set. This is followed by taking logarithm, and then DCT to obtain static coefficients appending along with Δ and ΔΔ in order to obtain higher-dimensional feature vector.

7.4.2.3 Pattern Classifier

In this study, Gaussian Mixture Model (GMM) is used as a two-class pattern classifier, where the two classes corresponds to the speech samples of the normal *vs.* whispered speech. The individual GMM is trained for each class using LFCC, MFCC, and TECC feature sets. The Expectation Maximization (EM) algorithm is used to find parameters of each GMM through an iterative optimization procedure. The log-likelihood (*llk*) score $s(X)$ for each test sample is estimated using the trained GMM as in:

$$s(X) = llk(X|\lambda_w) - llk(X|\lambda_n), \tag{7.8}$$

where λ_w, and λ_n represents the GMM trained on whisper, and normal speech samples, respectively, and X represents a new testing sample. The scores obtained helps to classify whether the unknown sample belongs to the natural or whisper. The robustness and feature discrimination power of our proposed feature set is also evaluated using Matthew Correlation Coefficient (MCC), F-measure, and J-statistics. Furthermore, we used a standard evaluation metric, i.e., Equal Error

Rate (EER) which is indicated on the Detection Error Trade-off (DET) curve for the whisper speech detection system [217]. The DET curve is used to study the performance of the SSD system. When operating point in the DET curve of False Acceptance Rate (FAR), and False Rejection Rate (FRR) or miss probability is *equal*, then it is referred to as EER.

7.4.3 Experimental Results

The experiments are performed on wTIMIT and CHAINS corpus with TECC feature set are shown in Table 7.14. We observed the effect of two different frequency scales, namely, linear and Mel scale in the Gabor filterbank in order to obtain the subband filtered signals according to the center frequencies. When the features are extracted using linear frequency scale, the accuracy of the whisper speech detection was not high as when the features were extracted using Mel frequency scale for both the corpora. Along with observing the effect of frequency scale, we also observed the effect of applying the CMN technique for both the corpora. It can be observed from the Table 7.14 that the accuracy obtained from the Mel frequency scale along with CMN technique gave better accuracy of 92.22 % on wTIMIT, and 95.61 % on CHAINS corpus, respectively.

Table 7.14: Accuracy (in %) for WSD using TECC Feature Set on wTIMIT and CHAINS Corpora. After [22]

Corpus	CMN	Frequency Scale	Accuracy (%)
wTIMIT	×	Linear	84.65
	✓	Linear	90.93
	×	Mel	82.78
	✓	Mel	**92.22**
CHAINS	✓	Linear	86.88
	✓	Mel	**95.61**

In addition, we also observe the feature discrimination power using F measure, J-statistic, and MCC as shown in Table 7.15. It can be observed that the TECC feature set has high values for all the measures as compared to the other feature sets and thus, it is more discriminative to classify natural *vs.* whisper speech signals for both the corpora.

Furthermore, the performance evaluation metric is computed in terms of % EER for all the feature sets. It can be observed from the Table 7.16 that we obtain lower EER for the TECC feature set compared to the other MFCC and LFCC feature sets. For wTIMIT corpus, the low EER with TECC feature set is 6.69 % and for CHAINS corpus, it is 4.46 %. The performance evaluation is also shown in

Table 7.15: Analysis of Feature Discrimination Power using F-measure, J-statistic, and MCC. After [22]

Corpus	Feature Sets	MCC	F-measure	J-measure
wTIMIT	LFCC	0.73	0.89	0.75
	MFCC	0.61	0.83	0.63
	TECC	**0.83**	**0.93**	**0.86**
CHAINS	LFCC	0.67	0.83	0.67
	MFCC	0.43	0.64	0.44
	TECC	**0.92**	**0.95**	**0.91**

Table 7.16: Results in Terms of EER in (%) and Accuracy in (%). After [22]

Feature Sets	EER (in %)		Accuracy (in %)	
	wTIMIT	CHAINS	wTIMIT	CHAINS
LFCC	12.59	16.05	86.82	83.97
MFCC	17.37	5.97	80.12	94.06
TECC	**6.69**	**4.46**	**92.22**	**95.61**

Fig. 7.16 by the DET curves for MFCC, LFCC, and TECC feature sets. It can be observed that the miss probability of MFCC, and LFCC is very high for the given FAR, which is not a good case for whisper speech detection (WSD) system. There is a significant decrease in miss probability for TECC feature set for wTIMIT as shown in Figure 7.16(a). We observe similar pattern of results on CHAINS corpus, as shown in Figure 7.16(b). However, the TECC and MFCC feature sets have low miss probability, and LFCC feature set has very high miss probability.

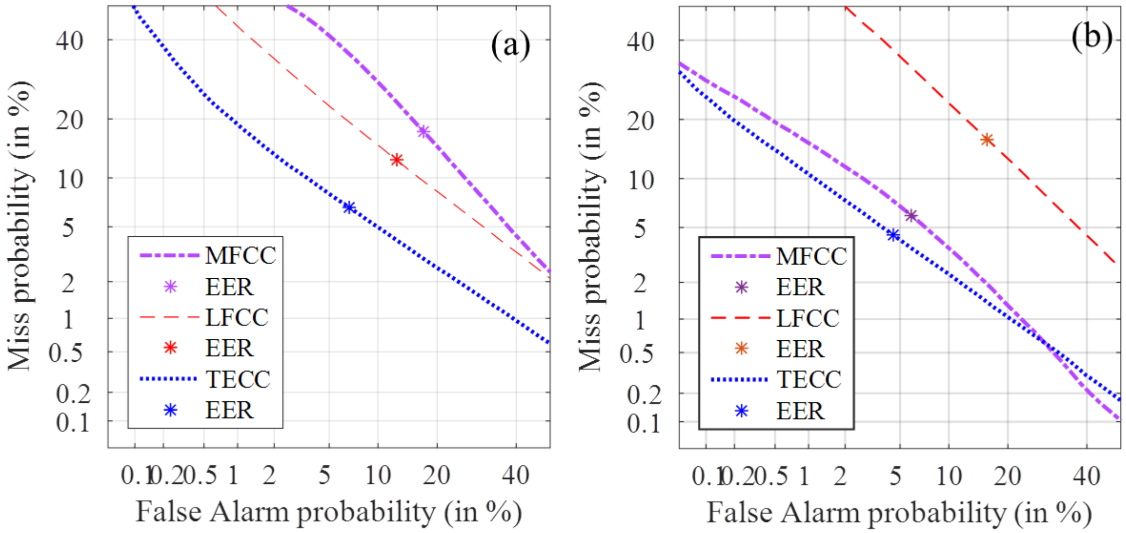


Figure 7.16: DET Curve for TECC, MFCC and, LFCC Feature Set for (a) wTIMIT, and (b) CHAINS Corpus. Arrow Indicates Relatively Best Performance of TECC Feature Set. After [22].

We also analyzed the trade-off between latency period *vs.* accuracy (as shown in Fig. 7.17) for (a) wTIMIT, and (b) CHAINS corpus. Here, latency period refers to the duration between the speech utterance produced to the system, and response from the system in terms % of accuracy. In the other words, if a system gives better accuracy for lower latency periods, then it means that this system would not be waiting for the entire utterance to judge whether the utterance is natural or whisper. Instead, lower levels of latency with higher accuracy would ensure that faster classification of natural *vs.* whisper utterances. In this graph, we considered frame-level accuracy. It can be clearly observed from the graph that the accuracy of TECC feature set increases continuously as the latency is increased. This behavior is expected because if the number of frames taking part in accuracy calculation increases, then the average value of accuracy tend to increase.

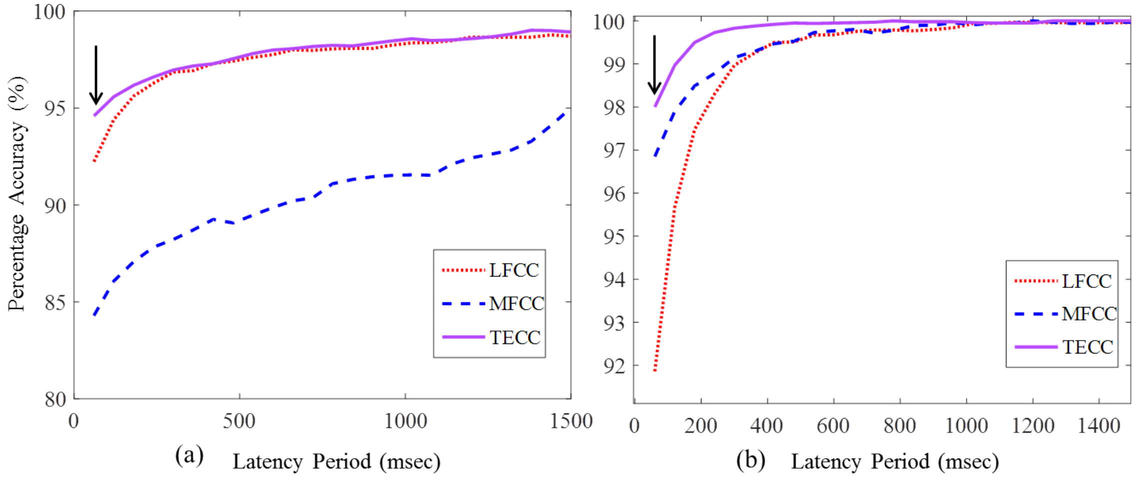


Figure 7.17: Accuracy (in %) *vs.* Latency Period for TECC, MFCC, and LFCC Feature Set for (a) wTIMIT, and (b) CHAINS Corpus. After [22].

In this Section, we explored TECC, MFCC, and LFCC feature set for normal *vs*. whisper speech classification. As whispered speech contains nonlinear and extremely turbulent airflow, the feature representation should incorporates both amplitude and frequency information of the signal. Hence, estimating the “true" total energy of the signal instead of estimating only kinetic energy into account of the signal. By listening to the speech samples of natural and whisper speech, it has been observed that the initial and the end portion of the utterance consists of silence regions. These silence regions produces ambiguity to the classification architecture. We tried to eliminate these regions. However, it is not the straightforward to remove such silent regions for the whispered signal because the amplitude of acoustic noise is much higher than the amplitude of the whisper component. Hence, noise cannot be directly removed from the whispered speech. Thus, developing efficient Voice Activity Detection (VAD) algorithms in whispered speech is also a potential area of research. With this development, we can improve the frame-level classification accuracy for the whispered speech and also higher accuracy at low latency period.

7.5 Acoustic Scene Classification (ASC)

Acoustic Scene Classification (ASC) is a challenging research problem which is seen as a subset of Computational Auditory Scene Analysis (CASA) [262]. It is the task of classifying acoustical scenes and events from the surrounding noise, silence, etc. where the environment can be a busy street, quite park, etc. Sounds carry information about our everyday environment, and events that happens around us. With recent advancements in technology especially in the field of machine learning, developing methods to capture this information can be invaluable to number of applications, such as searching for multimedia-based on audio content [263], designing automated cars, robots that depends on the context [264], intelligent monitoring systems to recognize activities using acoustic information and many more. It may look trivial for humans classifying an acoustical scene after hearing the audio sample. However, it is challenging to develop artificial systems that classify the acoustical scenes especially the scenes with sound sources in real-life environments, where often multiple sounds are present.

The schematic representation of Acoustic Scene Classification (ASC) task is shown in Fig. 7.18. In ASC task, audio signals recorded in different acoustical environments are used for the training of models. This model training uses the front-end features which can be in cepstral-domain or the log-energy coefficients that are obtained from the filterbank energies. Depending on the features, the models are prepared on the training data with the traditional or neural network-based classifiers. For the given test signal, the features are extracted and depending on the probabilities obtained from the trained models, the test signal is classified into corresponding acoustical scene. The first DCASE challenge was organized in 2013 to emphasize the problem of developing machines to perform ASC and to provide a publicly available database containing non-speech and non-music audio samples. This was followed by three more challenges in 2016, 2017, and 2018, where several researchers proposed various models using different classifiers, such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVM), tree-bagger classifiers along with audio features. Convolutional Neural Networks (CNNs) proved to be successful for diverse audio-related tasks, such as speech recognition, environmental sound classification, robust audio event recognition and thus,

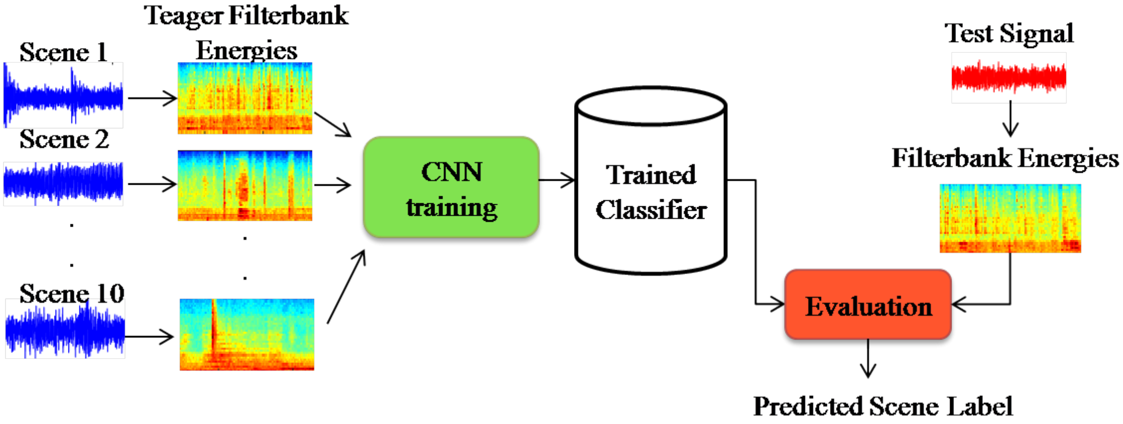


Figure 7.18: Schematic representation of proposed Acoustic Scene Classification (ASC) system. After [23].

motivated researchers to use these networks for acoustical scene classification as well [265].

7.5.1 Front-end Features

In this Section, we discuss the front-end features used for ASC task. The organizers of the DCASE 2018 challenge provided the baseline system that includes the Mel filterbank energies along with CNN classifier. We are comparing the Teager energy-based log-energy coefficients with the baseline system. We choose linear scale over the Mel scale as our experimental results and the analysis of spectral energy obtained from the Teager energy shows its better performance compared to the Mel scale.

Fig. 7.19 shows the analysis of (a) linear, and (b) Mel frequency scale used in Gabor filterbank. With linear frequency scale, it can be observed that the subband filters are equally distributed across the frequency range. Whereas, with Mel scale, initial filters are compressed at the lower frequencies and has expanded bandwidth with a few number of subband filters in higher frequency regions. This can be also observed from the placing of center frequencies for linear and Mel scale, where the linear scale has the center frequency varying linearly from 0 to 8000 Hz. On the other hand, for Mel scale, approximately 20 to 25 subband filters are covered within 2000 Hz frequency while the remaining 20 subband filters are placed between the frequency range of 2000-4000 Hz. The spectral energy obtained from the linear scale with Teager energy-based approach shows the differences in lower as well as higher frequency regions compared to the Mel scale as shown in the last panel of Fig. 7.19.

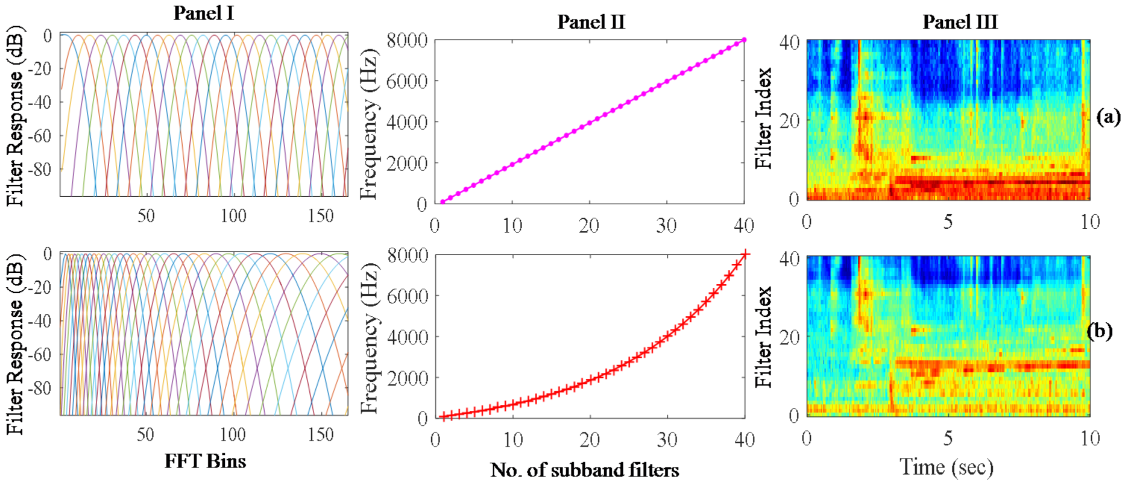


Figure 7.19: Filterbank Analysis of Acoustical Scene: (a) Linear Scale *vs.* (b) Mel Scale. Panel I: Filterbank Response, Panel II: Frequency Scale, and Panel III: Filterbank Energies.

7.5.2 Experimental Setup

7.5.2.1 Database

DCASE 2018 challenge provides the audio signal data for five different tasks including acoustic sound classification, audio tagging, bird audio detection, sound event detection in domestic environments using weakly-labeled data, and monitoring domestic activities based on multi-channel acoustics. In particular, for Task-1 (ASC), based on the data preparation, there are three different sub-challenges. For Task-1A, the devices used for recording of development and evaluation data are the same, while for the other two evaluations, data can be recorded using other devices which are not used in development set. In this chapter, we focused on Task-1A sub-challenge, i.e., ASC with pre-defined classes.

The DCASE 2018 challenge focuses on detection and classification of acoustic scenes and events, in which there are total of 5 sub-challenges. In this chapter, we are focusing on acoustic sound classification (Task-1). Based on the recording devices for development and evaluation sets, the Task-1 has three sub-challenges. The Task-1A challenge uses the same recording devices for both development and evaluation sets. The newly recorded TUT Urban Acoustic Scenes 2018 dataset is the largest freely available dataset that consists of ten different acoustic scenes, such as airport, park, metro station, etc. This dataset consists of 24 hours of high quality audio that was recorded in six European cities making it relatively much harder than the previous datasets due to its high acoustical variability. It is the first dataset containing data recorded in multiple countries, in addition to the

Table 7.17: Distribution of TUT Urban Acoustic Scenes 2018 Development Data into Train and Test Subset. After [23]

Scene Label	Development Set	Train Subset	Test Subset
Airport	864	599	265
Bus	864	622	242
Metro	864	603	261
Metro station	864	605	259
Park	864	622	242
Public square	864	648	216
Shopping mall	864	585	279
Street, pedestrian	864	617	247
Street, traffic	864	618	246
Tram	864	603	261
Total	8640	6122	2518

data recorded with mobile devices. The total development set consists of 8640 segments of 10 seconds audio signal, i.e., 864 segments for each acoustic scene. The development set is further sub-divided into two parts, namely, train and test subsets. The baseline system includes convolutional neural network (CNN) with log-Mel filterbank energies as features, and the recommended cross-validation setup for determining the performance on the sub-tasks.

7.5.2.2 Convolutional Neural Network (CNN)

The challenge organizers provided a baseline system, where they used CNN as the classifier with Mel filterbank energies. This baseline CNN architecture is based on one of the top-ranked submission from DCASE 2016 [266], where changes are made to the regularizer used and the number of layers in the network. The baseline architecture consists of 2 convolutional layers, and 1 Fully-Connected (FC) layer. The input layer is of size 40×500, where 40 represents the number of subbands, and 500 denotes the number of frames. The first convolutional layer consists 32 subband filters, and 7×7 stride size, followed by batch normalization with Rectified Linear Unit (ReLU) as the activation function. The output obtained is passed through 2-D max-pool layer with stride size chosen as 5×5. The second CNN layer performs almost the same operation with changes made to the number of subband filters and pool size as 64, 4×100, respectively. Later, the output is flattened to give input to the dense layer with 100 neurons, and ReLU activation. Finally, in order to obtain the probabilities, softmax is applied on the dense layer.

In this Chapter, for ASC task, we trained the filterbank Teager energies on the

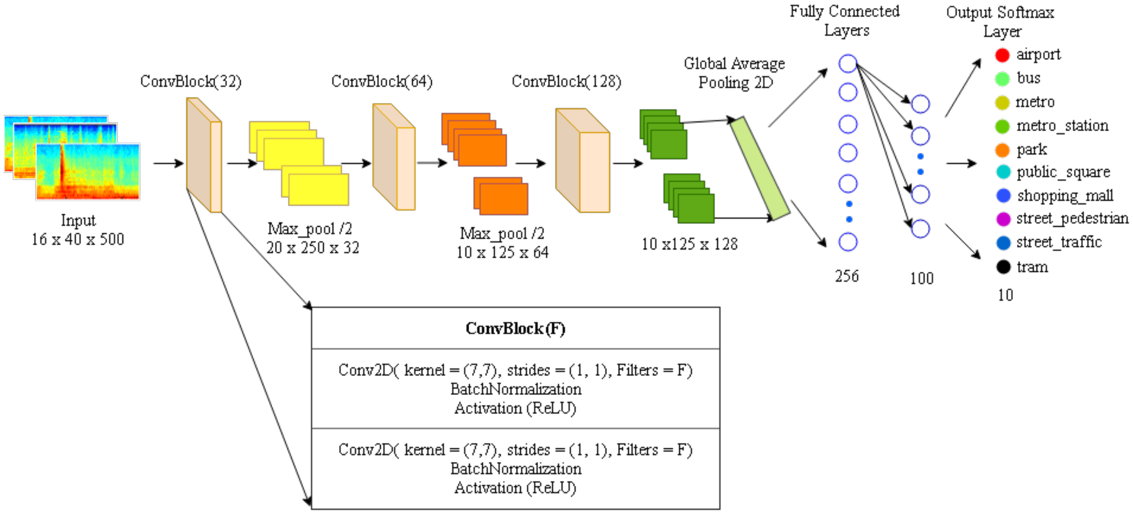


Figure 7.20: Architecture of Convolutional Neural Network (CNN) Used for ASC Task. After [23].

CNN classifier as shown in Fig. 7.20. The CNN architecture used consists of a series of Convolutional Block (ConvBlock), and pooling layers followed by FC layers. A ConvBlock is a combination of two convolutional layers with 7 × 7 kernel size, and ReLU activation. A total of three ConvBlock with 32, 64, and 128 subband filters, respectively, are used in the architecture. The max-pooling layers with 2 × 2 kernel, and stride size are used in between the ConvBlocks in order to capture the important discriminative information. After the third ConvBlock, a global average pooling layer is used to pool the data across the filter maps. On the pooled data, two FC layers are used with 256 and 100 neurons, respectively. Finally, an output softmax layer with 10 neurons indicating the acoustical scenes is used to classify the input data. The batch normalization and dropout layers are used as the regularization parameters in the model. Batch normalization layers are used in between the convolutional layer, and its activation functions. The batch normalization across the filter maps helps to improve the generalization of the network [267]. A dropout of 0.3 was used after every pooling and FC layers. Adam optimizer with 0.001 learning rate is used to train the network with batch size of 16 for 200 epochs. The epoch with the best accuracy on test set is considered.

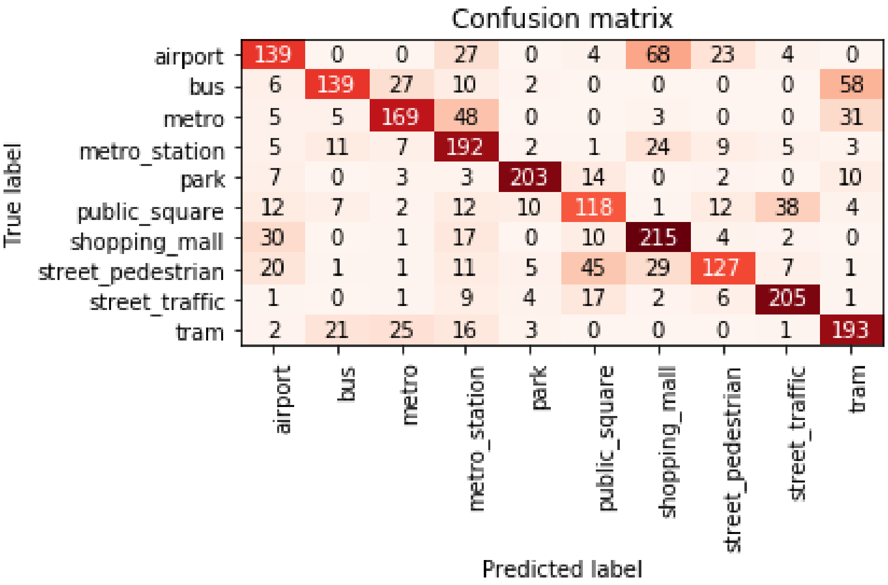
7.5.3 Experimental Results

The performance of the Teager energy-based filterbank features is shown in Table 7.18 in terms of accuracy (in %) on the test set. In particular, we compare the performance of baseline and Teager filterbank energies obtained with linear and Mel frequency scales in Gabor filterbank. It can be observed that the accuracy obtained from the linear frequency scale gave better results compared to the Mel frequency scale (except for audio signals recorded in the street acoustic scene). The baseline system gave an average accuracy of 59.7 %, with classwise results varying from 40.4 % to 80.5 %. Teager energy-based subband features that are obtained from Mel and linear scales gave 64 % and 67.3 %, respectively, have the classwise results from 44 % to 85 % (for Mel scale) and 51.4 % to 83.9 % (for linear scale). For individual acoustical scene classification, we have shown the confusion matrices in Fig. 7.21 for the proposed features with linear frequency scale used in Gabor filterbank. It can be observed from confusion matrix that with linear frequency scale, we obtained the better performance compared to the Mel frequency scale. For the street pedestrian scene, the Mel frequency scale performed better with less ambiguity than with linear scale. However, for rest of the acoustical scenes, linear frequency scale gave better performance. Comparing the system performance with linear scale, for most of the similar scenes, it gave same performance, such as for park and street traffic approximately 83 % accuracy and metro station and tram gave 74 % accuracy. The most difficult task is to detect the scenes of airport and street pedestrian giving lowest performance of 52.5 % and 51.4 %, respectively. This is also observed in terms of confusion matrix (as shown in Fig. 7.21), where we can observe more ambiguity for airport, metro, and street pedestrian classes.

Table 7.18: Individual Scene Accuracy (%) on Test Set for Acoustic Scene Classification, Subtask A in DCASE 2018 Challenge. After [23].

Scene Label	**Baseline**	**Teager-Energy based Subband Features**	
		Mel Scale	**Linear Scale**
Airport	72.9	49.4	52.5
Bus	62.9	51.2	57.4
Metro	51.2	49.4	64.8
Metro Station	55.4	71.4	74.1
Park	79.1	78.5	83.9
Public square	40.4	44.0	54.6
Shopping mall	49.6	77.1	77.1
Street, pedestrian	50.0	61.5	51.4
Street, traffic	80.5	85.0	83.3
Tram	55.1	72.0	73.9
Average	59.7	64.0	**67.3**

In this Section, we proposed Teager energy-based log-filterbank energies for acoustical scene classification task. We analyzed the differences of the audio signals when recorded in different acoustical environments, such as airport, bus, tram, metro, etc. The filterbank Teager energies obtained from the linearly-spaced



True label \ Predicted label	airport	bus	metro	metro_station	park	public_square	shopping_mall	street_pedestrian	street_traffic	tram
airport	139	0	0	27	0	4	68	23	4	0
bus	6	139	27	10	2	0	0	0	0	58
metro	5	5	169	48	0	0	3	0	0	31
metro_station	5	11	7	192	2	1	24	9	5	3
park	7	0	3	3	203	14	0	2	0	10
public_square	12	7	2	12	10	118	1	12	38	4
shopping_mall	30	0	1	17	0	10	215	4	2	0
street_pedestrian	20	1	1	11	5	45	29	127	7	1
street_traffic	1	0	1	9	4	17	2	6	205	1
tram	2	21	25	16	3	0	0	0	1	193

Figure 7.21: Confusion Matrix w.r.t Performance of Teager Energy-Based Subband Filters with Linear Frequency Scale for Task 1A Performance. After [23].

Gabor filterbank gave high spectral energy density compared to the traditional and Mel scale filterbank energies. In addition, we also observed that the performance of the similar scenes gave approximately the same accuracy. For the acoustic scenes, such as airport and street pedestrians performance is low, which can be observed through the ambiguity in detection.

7.6 Chapter Summary

The application of Teager energy-based feature set on audio classification tasks, namely, the ASR, VCS, WSD along with ASC task are presented in this Chapter. The experiments on the ASR task shown the improved performance by GTFB spectral features on near and far-field corpus. Significant reduction in % WER was achieved using the system combination using MBR decoding of the MFCC and GTFB-based features. The experiments on the replay SSD task on VAs demonstrate that it performs better than the baseline system with CQCC feature and GMM as pattern classifier. In addition, the experiments on the WSD with TECC feature set performed better compared to MFCC feature set. Finally, the experiments on the ASC show that the proposed Teager energy-based spectral features performed well compared to the MFCC-based baseline. In the next chapter, we summarize the entire thesis and present some of the limitations of work presented in thesis and potential future research directions.

CHAPTER 8

Summary and Conclusions

In this chapter, a summary of this thesis work is presented along with the limitation of the current work and future research directions.

8.1 Summary of Work Presented in the Thesis

The following is a summary of the research work done in the entire thesis:

- In this thesis work, various signal processing-based feature sets, such as TECC, ESA-IFCC, ESA-IACC, VTECC, VESA-IFCC, VESA-IACC, and AWFCC are presented. The feature sets are based on Teager Energy Operator (TEO) and Energy Separation Algorithm (ESA) using Gabor filterbank to obtain narrowband filtered signals. Compared to the earlier studies using TEO, we explored the TEO-based features for the Spoof Speech Detection (SSD) tasks for ASV and VAs. The feature sets are successfully applied for various other speech and audio processing applications as well. The motivation behind using TEO and ESA demodulation feature was as follows:
  - TEO is known to capture property of airflow pattern in the vocal tract system during natural speech production and hence, exploit it for SSD task.
  - The ESA is used to develop narrowband filtered speech signals, which are modeled using AM-FM signals to account for time-varying amplitude envelope and instantaneous frequencies [57].
  - The ESA approach do not require the computationally complex task of phase unwrapping (as it is required for HT-based approach of analytic signal generation).
  - To estimate the IA and IF components with ESA approach, only five consecutive samples samples are required.

  - Significance of extracting proposed feature sets with and without integrating the filterbank is investigated.
  - The slow and fast-varying temporal modulations obtained at different time scales have the distortion for the replay speech signal compared to the natural speech.
  - The IF component estimated for the subband filtered signal shows the damping in the fluctuation for the replay signal around the center (carrier) frequency.
  - For the same time scale from where the IF fluctuation started having tilt from its center frequency, *sinc*-like patterns are observed in replay signal than its natural counterpart in the voiced regions.
  - The spectral energy density obtained from the Teager Energy Operator (TEO), shows the difference for the natural, and its corresponding replay speech signal in all the frequency regions, which is not captured by the traditional spectrogram.
  - Relative significance of applying the CMN (highpass filtering) *vs.* CMVN (adaptive gain control) method (which is originally analyzed for robust speaker recognition task) is analyzed.

- The background studies required to understand the SSD task and applications were discussed in Chapter 2. The detailed architecture of the TEO and the mathematical derivation to capture reverberation during replay mechanism were presented in Chapter 3. The Teager energy traces obtained are distinct for different acoustical environments. The features extracted from the TEO along with Gabor filterbank are applied several standard spoofing databases in the SSD task. The proposed feature set performed very well compared to the corresponding baseline systems.

- To further improve the performance of the SSD, we explore ESA-based feature sets, namely, ESA-IFCC and ESA-IACC in Chapter 4. The experiments performed on the IA and IF-based feature sets presented using several evaluation factors, such as shape of subband filters, frequency scales, and the number of subband filters. Furthermore, in Chapter 5, we extend the work using generalized TEO, i.e., by varying the samples of past and future signal with a constant arbitrary integer know as *lag parameter or Dependency Index (DI)*. We investigate the advantage of VESA over ESA by varying the DI to capture the *hidden* dependencies, and dynamics in the sequence of samples of speech signal.

- Furthermore, in Chapter 6, we discussed about the importance of using the combined information of IA and IF components for replay SSD task. The IA components estimated from the ESA technique is severely affected by the noise and multipath interference (due to replay mechanism) this noise is explored by the IF components. The significance of using IA and IF components performed better compared to the baseline system on the ASVspoof 2017 spoofing database.

- After successful application for SSD task for ASV using Teager energy-based feature sets, we have also applied TEO-based feature sets, for other applications. To show the capability of proposed feature set, we explore the use of Teager energy spectral features-based acoustic model for near-field *vs.* far-field ASR tasks, where the GTFB feature set was extracted from Mel-spaced Gabor filterbank. The TEO preserves the amplitude and frequency modulation of a resonant signal, and it improves the time-frequency resolution. The noise suppression capability of TEO indeed helps for robust ASR task. The experiments are performed on both LibriSpeech (near-field) and CHiME-3 (far-field) corpus. Significant reduction in % WER was achieved using the system combination using MBR decoding of MFCC, and GTFB-based features. The next application we explore is to develop countermeasures for replay SSD task for Voice Assistant (VAs) task on the ReMASC corpus. The proposed features perform significantly better compared to the CQCC feature sets using GMM as pattern classifier. In addition, we also explored the Teager energy-based feature set for whisper speech Detection (WSD) task on the wTIMIT and CHAINS corpus. Furthermore, the Acoustic Scene Classification (ASC) task on DCASE 2018 challenge database using the Teager energy-based spectral features is explored to classify different acoustic scenes. The performance of the proposed feature set gave better accuracy compared to the MFCC feature set using CNN as classifier.

8.2 Limitations of the Current Work

Our proposed feature sets are one of the contributions towards the research in the Spoof Speech Detection SSD task. However, there are certain limitations of our proposed feature sets as described below:

- The amount of reverberation might be even more in some of the bonafide far field samples compared to near-field high quality replay speech and thus, directly relying on the amount of reverberation to test the replay spoof is little risky in this context, and hence, a more detailed study and generalized countermeasure is required to overcome the replay detection task.

- Our hypothesis using TEO to capture characteristics of reverberation due to replay, however, this may not be case for replay attack in outdoor (non-room) acoustic environment. Hence, our method is not likely to produce good results for replay recorded in outdoor environment.

- In addition, our approach ignore the fact bonafide utterances might also contains reverberant noises, such as in smart speakers.

- For high-level threat and high quality devices used during playback, and recording, the EERs are quite high. This needs further investigations to detect the high-level replay configuration threat.

8.3 Future Research Directions

Future research directions include possible solutions to the above mentioned limitations and further advancements for SSD task as described below:

- **Analysis of Reverberation using AM-FM Approach:** To investigate further the effect of reverberation on AM-FM components of a signal and its relation with the TEO framework for the different acoustic environments and intermediate device conditions on the replay speech.

- **Performance of Joint Protocol of SSD system with ASV Systems**: The current studies of countermeasures and ASV systems are carried out separately. What user would like to have is a secure and accurate ASV system. However, a more robust ASV system to noise and channel variations may become less secure against spoofing attacks. As there is no guarantee of having a better performing countermeasure that provides lower EER and also reliable for the ASV system performance. Hence, with the progress made in the research of spoofing detection, evaluation metrics must evolve to reflect the joint protocol system performance.

  Recently, the study reported in [30, 106] proposed a tandem Detection Cost Function (t-DCF) metric. It is an elegant solution to the assessment of combined spoofing countermeasures and ASV system. One of the initial attempts in this direction was reported in the mimic spoof detection task [268].

The t-DCF is used with the assessment of ASV systems that are combined with spoofing countermeasures (CM) as shown in Fig. 8.1 [30, 106].

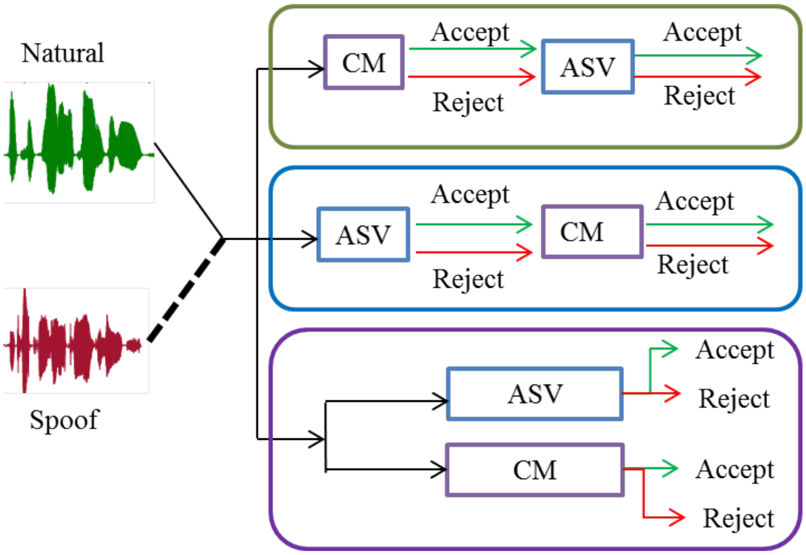


Figure 8.1: Performance assessment of combined ASV system and countermeasure (CM) that are combined as (top): CM followed by ASV, (middle): ASV followed by CM and (end): parallel both CM and ASV. Adapted from [30].

The perfect countermeasure system has an EER of 0 %. In particular, when the position of the miss (or false rejection), and the false alarm (or false acceptance) rate of the countermeasure (CM), i.e., P_{miss}^{cm}= P_{fa}^{cm}=0.

$$\text{t-DCF}_{CM(\theta)} = C_{miss}^{asv} \cdot \pi_{tar} \cdot P_{miss}^{asv}(\theta) \tag{8.1}$$

$$+C_{fa}^{asv} \cdot \pi_{non} \cdot P_{fa}^{asv}(\theta), \tag{8.2}$$

where C_{miss}^{asv} is the cost of ASV system rejecting a target trial, C_{fa}^{asv} is the cost of ASV system accepting a non-target trial, $P_{miss}^{asv}(\theta)$ and $P_{fa}^{asv}(\theta)$ are the position to define FRR rate, and the FAR of the ASV system at threshold (θ).

- **Joint Protocol with SSD system and VAs Systems**: The task of ASV and VAs though look similar, however, they have some important differences, such as, different user acoustic scenarios. The VAs use the far-field speech recognition with variety of different acoustic environment. In addition, the VAs use the multi-microphone array and ASV in general uses single array. This gap between the ASV and VAs degrades the performances by resulting in high EER. Hence, this grabs the major attention for the researchers to develop an algorithm for the joint protocol of SSD and VAs system performance.

- **Liveness Detection**: The use of high-quality recording loudspeaker or playback device to record/playback the speech signal. In this process, the quality of signal captured becomes indistinguishable from live human voice. This high quality device makes the speech signal impossible to detect that depends on the acoustic cues. This gives rise to investigate further on the liveness detection of human voice.

- **Signal Degradation Conditions**: Current publicly available spoofing databases are developed in clean conditions. However, the recent replay database was recorded under various acoustic environmental conditions. For ASVspoof 2015 challenge database, the noisy database was developed by adding various noises at different Signal-to-Noise Ratio (SNR) levels. Further investigations are required as to how the diversity of different noise types affects the SSD performance. In addition, the study is required to observe the effect on SSD, when the additive noise is added manually, and when the noise is added naturally via the acoustic environment. Hence, the countermeasures must be developed that it should be robust to signal degradation conditions as well.

- **Robustness in ASV Implies Vulnerability**:
  In practice, we would like an ASV system to be robust against variations, such as microphone and transmission channel, intersession, acoustic noise, speaker aging, etc. A robust ASV system may become vulnerable to various spoofing attacks as it tries to nullify these effects, and normalize the spoofing speech towards the natural speech. Thus, robustness and anti-spoofing security should be addressed separately. It is worth to study how features, classifiers, and systems are designed to be both robust and secure.

Appendix A. Performance Measures

A.1 % Equal Error Rate (EER)

A detection task or classification task can also be viewed as involving a trade-off between the two types of errors, namely, miss detection and false alarm. The miss detection or the False Rejection Rate (FRR) is the probability that the classifier fails to detect a match between the input pattern, and a matching class in the database [31]. FRR measures the percentage of valid inputs that are incorrectly rejected in the classification task. The false alarm or the False Acceptance Rate (FAR) is the probability that the classifier incorrectly matches the input pattern to a non-matching class in the database [31]. FAR measures the percent of invalid inputs that are incorrectly accepted in the classification task. A detection error trade-off (DET) graph is a graphical plot of error rates for binary classification systems, plotting the FRR *vs.* FAR [31]. Since FAR, and FRR are opposite functions (when one monotonically increases, the other monotonically decreases and vice-versa), there is a trade-off in the error reduction in the detection task and hence, the name DET curve. The point where FRR and FAR are equal is called as the % Equal Error Rate (EER), which is generally used as the performance measure. An example of the DET curve is shown in Figure A.1 for the SSD task. A lower FAR means higher security against spoof speech, i.e., a desirable attribute for spoof-resistant ASV system. A lower FRR means higher convenience of the system performance.

A.2 Half Total Error Rate (HTER)

The performance evaluation metrics for BTAS 2016 database are considered according to the protocol used in the BTAS 2016 speaker anti-spoofing challenge [100]. The results on the development set are reported in terms of EER, and on the test data in terms of Half Total Error Rate (HTER). The evaluation of the replay attack systems was done based on the *false rejection rate* (FRR), and *false acceptance rate* (FAR), that in turn depends upon a threshold, θ. We use the development set

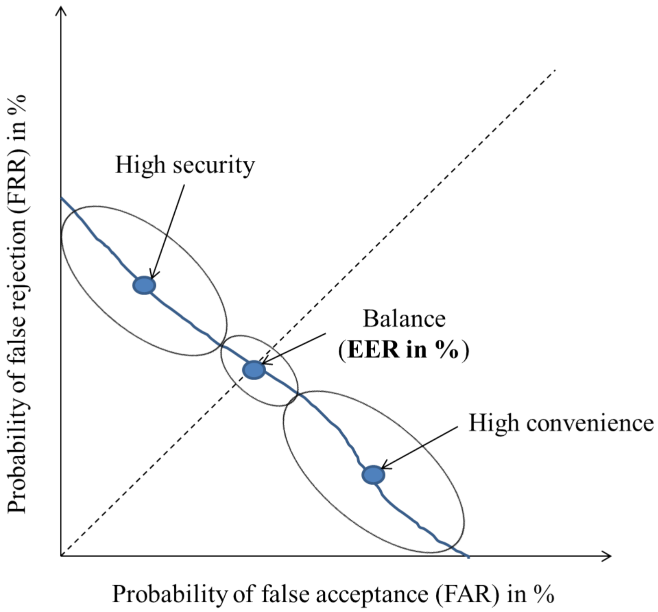


Figure A.1: An Example of the DET Curve. After [31].

to determine threshold, θ_{dev}. The evaluation performance of the system is then computed as the HTER:

$$\theta_{dev} = \arg\min_{\theta} \frac{\mathrm{FAR}_{dev}(\theta) + \mathrm{FRR}_{dev}(\theta)}{2}, \tag{A.1}$$

$$\mathrm{HTER}_{eval}(\theta) = \frac{\mathrm{FAR}_{eval}(\theta_{dev}) + \mathrm{FRR}_{eval}(\theta_{dev})}{2}. \tag{A.2}$$

A.3 % Classification Accuracy

The performance of the classification task is measured by the classification accuracy. If $\hat{z}_i$ is the predicted value of the i^{th} sample, and z_i is the corresponding true value, then the % classification accuracy (the fraction of correct prediction) over a total of N samples is defined as [269]:

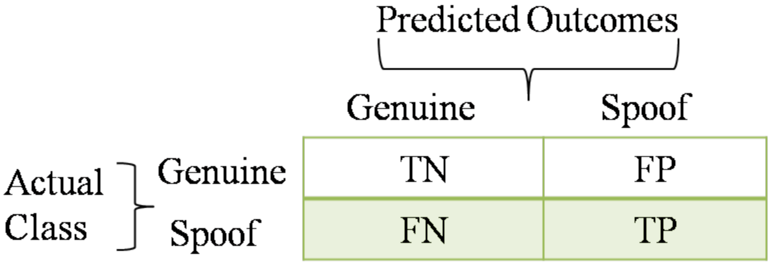
$$\text{\% Classification Accuracy} = \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{I}(\hat{z}_i = z_i) \times 100, \tag{A.3}$$

where $\mathbb{I}(\cdot)$ is an indicator function with $\mathbb{I}$=1, when $\hat{z}_i = z_i$, otherwise $\mathbb{I}$=0.

A.4 Performance Measures from Confusion Matrix

The confusion matrix of a binary classification task shows how errors are distributed across the classes [270]. The example of a confusion matrix for a classification task is shown in Figure A.2 for genuine *vs.* spoof speech. The rows indicate the actual classes, and columns indicate the predicted outcome of the pattern classifier [270]. Since our task is to detect the spoof speech, we denote the results associated with spoof class as positive, and with genuine class as negative. Given the labels of actual and predicted classes by the classifier, there are four outcomes possible [270]:

- True Positive (TP): Actual class is spoof and predicted spoof
- True Negative (TN): Actual class is genuine and predicted genuine
- False Positive (FP): Actual class is genuine and predicted spoof
- False Negative (FN): Actual class is spoof and predicted genuine



		Predicted Outcomes	
		Genuine	Spoof
Actual Class	Genuine	TN	FP
	Spoof	FN	TP

Figure A.2: The Details of a Confusion Matrix for the Binary Classification Task.

In the case of *k*-fold CV, we find the combined confusion matrix (i.e., all the entries in the matrix are summed for all the folds). Various other performance measures can be obtained from the confusion matrix. The numbers along the major diagonal indicates (TP and TN) the correct decisions made by the classifier [270]. The classification accuracy can also be obtained from TP, TN, and a total number of instance of both the classes (i.e., P+N) as follows [270]:

$$\text{Classification Accuracy } (in\%) = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}. \tag{A.4}$$

Another important performance measure is the F1-score, also known as F-measure. The range of F-measure is between 1 and 0, where 1 represents the perfect prediction and 0 means the worst. The F-measure is defined as follows [270]:

$$\text{F-Measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{A.5}$$

The F-measure does not take TN into account. Hence, we also used another performance measure called Youden's J-statistic or informedness [271]. The range of the J-statistic is between -1 and +1, where -1 indicates no agreement between the observation and the prediction, and +1 represents a perfect prediction. The J-Statistic estimates the probability of an informed decision and is given by [271]:

$$\text{J-Statistic} = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1. \tag{A.6}$$

Another important performance measure is the Matthews Correlation Coefficient (MCC) [272]. It takes into account TP, TN, FP, FN, and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. The range of MCC is between -1 and +1, where +1 indicates a perfect prediction, 0 means no better than just a random prediction, and -1 indicates a total disagreement between the observation and the prediction. MCC is expressed as [272]:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TP} + \text{FN})(\text{TN} + \text{FP})}}. \tag{A.7}$$

A.5 % Word Error Rate (WER)

The standard performance metric for Automatic Speech Recognition (ASR) systems is the Word Error Rate (WER) [273]. The WER is computed for the decoded word sequence in the ASR output against the reference transcription. The % WER is defined as follows [273]:

$$WER = \frac{S + D + I}{N} \times 100, \tag{A.8}$$

where
S = Number of substitutions (one word is replaced with another one),
D = Number of deletions (word is missed out),
I = Number of insertions (word is added),
N = Total number of words in the reference transcription.

In the case of the phone recognition task, the reference is the phonetic transcription (not at the word-level), and the ASR decoder also produces phone sequences in the output. In such a case, the same performance measure is applied, however, instead of words, we use phones. Hence, it is also called the % Phone Error Rate (PER).

Appendix B. Energy Separation Algorithm (ESA)

In this Section, an alternative DESA that avoids the previous half sample shifts in the estimated frequency signal by using a symmetric difference to approximate the first derivative of $x(t)$ is discussed [57,190]. Consider first a discrete-time AM-FM signal $x(n) = a(n)cos[\phi(n)]$ whose instantaneous frequency signal $\Omega_i(n)$ is a finite sum of cosines. It's symmetric difference is:

$$s(n) = [x(n+1) - x(n-1)]/2, \tag{B.1}$$

$$s(n) = D(n) + E(n), \tag{B.2}$$

where

$$D(n) = a(n)cos\phi + (n+1) - cos\phi + (n-l)l/2, \tag{B.3}$$

$$E(n) = [a(n+1) - a(n)]cos\phi + (n+1)]/2 + [a(n - a(n-l)]cos[\phi + (n-1)]/2. \tag{B.4}$$

since, $\Omega_f << 1$,

$$c(n) = -sin\left[\frac{\phi + (n+1) - \phi(n-1)}{2}\right] \cdot sin\left[\frac{\phi + (n+1) + \phi(n-1)}{2}\right], \tag{B.5}$$

$$c(n) \approx -sin[\Omega_i(n)]sin[\phi(n)]. \tag{B.6}$$

Now $D_{max} \approx a_{max}sin(\Omega_i)_{max}$ and $E_{max} \approx 2a_{max}sin(\Omega_a/2)$. Hence, the order of magnitude of D is much larger than that of E. Thus, ignoring E,

$$s(n) \approx -a(n)sin[\Omega_i(n)_{]}sin[\phi(n)]. \tag{B.7}$$

Since $O_m << Q_c$, the amplitude $a(n)sin(\phi_i(n))$ of $s(n)$ has an effective bandwidth of $\Omega_a + \Omega_f$. Hence,

$$\Psi[s(n)] = a^2(n) sin^4 \Omega_i(n) \tag{B.8}$$

The above analysis yields the following formulas for estimating the time-varying frequency and amplitude envelope of the AM-FM signal:

$$\Omega_i[n] \approx \frac{1}{2} arccos \left[1 - \frac{\Psi_d\{x_i[n+1] - x_i[n-1]\}}{2\Psi_d\{x_i[n]\}} \right], \tag{B.9}$$

$$a_i[n] \approx \frac{2\Psi_d\{x_i[n]\}}{\sqrt{\Psi_d\{x_i[n+1] - x_i[n-1]\}}}. \tag{B.10}$$

We call this the DESA-2 algorithm, where “2" implies the approximation of first-order derivatives by differences between samples whose time indices differ by 2. This DESA uses symmetric differences and thus, avoids having to involve values of Ω_i at non-integer time indices. The frequency estimation part assumes that $0 < \Omega_i(n)\pi/2$. Thus, the DESA-2 can be used to estimate instantaneous frequencies $< 1/4$ the sampling frequency. This does not present a problem because by doubling the sampling frequency it can be used to estimate frequencies up to $1/2$ the original sampling frequency. Note also that the formula with $arccos()$ can be replaced by an $arcsin()$ expression but this comes at the expense of an additional square-root operation per sample [57,190].

$$\Omega_i[n] \approx arcsin \sqrt{\frac{\Psi_d\{x_i[n+1] - x_i[n-1]\}}{4\Psi_d\{x_i[n]\}}}. \tag{B.11}$$

Appendix C. Noise Suppression Capability of TEO

The Teager Energy Operator (TEO) has the noise suppression capability first analyzed in [208] for speech recognition applications in the car noise scenario, for epoch estimation using various noises [274], and for person recognition in noisy environments [275]. Here, we discuss the noise suppression capability of TEO for the additive noise case. Let $x[n]$ and $\hat{x}[n] = x[n] + v[n]$ be clean, and noisy speech signal, where $v[n]$ is a zero-mean additive noise signal. The TEO profiles for $x[n]$, and $v[n]$ are given as:

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

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

The TEO profile for the noisy speech signal $\hat{x}[n]$ is calculated as:

$$\begin{aligned}\Psi\{\hat{x}[n]\} &= \hat{x}^2[n] - \hat{x}[n-1]\hat{x}[n+1], \\ &= (x[n]+v[n])^2 - (x[n-1]+v[n-1])(x[n+1]+v[n+1]), \\ &= x^2[n] + 2x[n]v[n] + v^2[n] - x[n-1]x[n+1] - x[n-1]v[n+1] \\ &\quad - v[n-1]x[n+1] - v[n-1]v[n+1].\end{aligned} \tag{C.3}$$

Rearranging the above terms and using Eq. (C.1) and Eq. (C.2), we get,

$$\Psi\{\hat{x}[n]\} = \Psi\{x[n]\} + \Psi\{v[n]\} + 2\hat{\Psi}\{x[n], v[n]\}, \tag{C.4}$$

where $\hat{\Psi}\{x[n], v[n]\}$ is called the cross-TEO between $x[n]$, and $v[n]$, which is given by:

$$\hat{\Psi}\{x[n], v[n]\} = x[n]v[n] - \frac{1}{2}x[n-1]v[n+1] - \frac{1}{2}x[n+1]v[n-1]. \tag{C.5}$$

Considering $x[n]$, and $v[n]$ as random variables, the expected value of the TEO is given as:

$$\mathbb{E}\left[\Psi\{\hat{x}[n]\}\right] = \mathbb{E}\left[\Psi\{x[n]\}\right] + \mathbb{E}\left[\Psi\{v[n]\}\right] + 2\mathbb{E}\left[\hat{\Psi}\{x[n], v[n]\}\right], \tag{C.6}$$

where $\mathbb{E}[\cdot]$ is an expectation operator. Since $v[n]$ is a zero-mean additive noise, and $x[n]$ and $v[n]$ are assumed to be statistically *independent* so that $\mathbb{E}\left[\hat{\Psi}\{x[n], v[n]\}\right] = 0$ and hence:

$$\begin{aligned}\mathbb{E}\left[\hat{\Psi}\{x[n], v[n]\}\right] = \mathbb{E}\left[x[n]v[n]\right] - \frac{1}{2}\mathbb{E}\left[x[n-1]v[n+1]\right] \\ - \frac{1}{2}\mathbb{E}\left[x[n+1]v[n-1]\right],\end{aligned} \tag{C.7}$$

Here, $\mathbb{E}\left[x[n]v[n]\right] = \mathbb{E}\left[x[n]\right]\mathbb{E}\left[v[n]\right] = 0$, since $\mathbb{E}[v[n]] = 0$, and similarly for other two terms in Eq. C.7. Hence, we have,

$$\mathbb{E}\left[\Psi\{\hat{x}[n]\}\right] = \mathbb{E}\left[\Psi\{x[n]\}\right] + \mathbb{E}\left[\Psi\{v[n]\}\right]. \tag{C.8}$$

The expected values in Eq. (C.8) can also be represented in terms of autocorrelation as follows:

$$\mathbb{E}\left[\Psi\{\hat{x}[n]\}\right] = R_{xx}(0) - R_{xx}(2) + R_{vv}(0) - R_{vv}(2), \tag{C.9}$$

where $R_{xx}(\tau) = \mathbb{E}[x[n]x[n-\tau]]$, and $R_{vv}(\tau) = \mathbb{E}[v[n]v[n-\tau]]$ are autocorrelation functions of clean and noise signals for lag τ, respectively. It is experimentally verified in [208] and [274] that, when the TEO is applied on the noise signal, $R_{vv}(0) - R_{vv}(2) \approx 0$. Hence, it can be proved that:

$$\mathbb{E}\left[\Psi\{\hat{x}[n]\}\right] \approx \mathbb{E}\left[\Psi\{x[n]\}\right]. \tag{C.10}$$

The Eq. (C.10) indicates that TEO when applied on the noisy signal, with the additive zero-mean noise, can suppress the noise and hence, TEO has the *noise suppression* capability.

Appendix D. ASR System Building in Kaldi

D.1 Data Preparation

- First get the wave and label files for both the training and testing set.
- Prepare wav.scp, spk2utt, utt2spk and text files for both the train and test set as follows:

```
ls *.wav | cut -d "." -f 1 >../cname.txt (instead of ls *.wav> ../cname.txt
because we do not need .wav extension)
ls -d $PWD/*.wav > ../cpath.txt
cd ..
paste cname.txt cpath.txt > wav.scp
paste cname.txt cname.txt > spk2utt
paste cname.txt cname.txt > utt2spk
ls *.lab | cut -d "." -f 1 > ../labname.txt
cat *.lab > ../labcont.txt
cd..
paste labname.txt labcont.txt >text
```

- In kaldi/egs/ make a chime3 directory In chime3 directory make versions of your system, e.g., s1, s2... etc. Here, we make s1 directory inside chime3 directory.
- Create a data directory in chime3. Make train and test folders inside data directory.
- Place the training and testing files wav.scp, spk2utt, utt2spk and text in data/train and data/test directory, respectively.

D.2 Language Model Preparation

- Create a folder named dict in the /s1/data/local. Put following files in this folder:lexicon.txt, nonsilence_phones.txt, optional_silence.txt, silence_phones.txt

- From the text of test set, find unique words and store them in words.txt

```
awk ’print $2’ < text > testwords.txt
sort -u testwords.txt > words.txt
```

- Use fst.sh and generate_bigram.py files. The content of the fst.sh is as follows:

```
# ./fst.sh contents
#!/bin/bash
. ./cmd.sh
. ./path.sh
main_dir=/home/daiict/kaldi/egs/chime3/s1/
data_dir=$main_dir/data
dict_dir=$main_dir/data/local/dict
tmp_dir=$main_dir/data/tmp
lang_dir=$main_dir/data/lang
mkdir -p $tmp_dir
utils/prepare_lang.sh $dict_dir ’!SIL’ $data_dir/local/actual_overal
$main_dir/data/lang || exit 1;

python local/generate_bigram.py $tmp_dir/words.txt
> $tmp_dir/wp_gram.txt

local/make_rm_lm.pl $tmp_dir/wp_gram.txt > $tmp_dir/G.txt

fstcompile --isymbols=$lang_dir/words.txt
--osymbols=$lang_dir/words.txt --keep_isymbols=false
--keep_osymbols=false $tmp_dir/G.txt > $lang_dir/G.fst
utils/validate_lang.pl $lang_dir
```

- generate_bigram.py program is used to generate the wp_gram from the district or commodity list. The content of this file is as follows:

```
#!/usr/bin/env python
import sys
from collections import defaultdict
word_list = [xx.strip().split() for xx in open(sys.argv[1])]
word_list = [ ["SENTENCE-END"] + xx + ["SENTENCE-END"] for xx in
word_list ]
suc_list = defaultdict(set)
for line in word_list:
for w1, w2 in zip(line[:-1], line[1:]):
suc_list[w1].add(w2)
list_of_keys = suc_list.keys()
list_of_keys.sort()
for ww in list_of_keys:
print ">" + ww
for ss in suc_list[ww]:
print " " + ss
```

- Change the paths in the fst.sh file and execute it; make sure to clear all the errors in this step. If it runs successfully, then you have no errors of mismatch in label files and lexicon. Do check G.fst file for binary file and not of very small size (in few bytes).

D.3 Feature Extraction

In this Section, we will extract the MFCC feature set that will be used to build the GMM-HMM systems, and the Mel filterbank (FBANK) feature set that will be used to build the hybrid DNN-HMM systems. Here, “nj 10" indicates the number of jobs to extract the features in parallel.

- The MFCC feature set is obtained as follows:

```
mfccdir=mfcc
for x in test train; do
    steps/make_mfcc.sh --cmd "$train_cmd" --nj 10 $datadir/$x
    exp/makemfcc/$x $mfccdir || exit 1;
    steps/compute_cmvn_stats.sh $datadir/$x exp/makemfcc/$x $mfc-
cdir || exit 1; done
```

- The FBANK feature set is obtained as follows:

```
fbankdir=fbank
for x in test train; do
    steps/make_fbank.sh --cmd "$train_cmd" --nj 10 $datadir/$x
    exp/makefbank/$x $fbankdir || exit 1;
    steps/compute_cmvn_stats.sh $datadir/$x exp/makefbank/$x
$fbankdir || exit 1; done
```

D.4 Acoustic Modeling GMM-HMM

In this Section, we will show how to build GMM-HMM system in KALDI.

- Monophone GMM-HMM system can be build by the following commands:

```
expdir=mono_mfcc
steps/train_mono.sh --nj "$train_nj" --cmd "$train_cmd" $datadir/train
data/lang exp/$expdir || exit 1;
utils/mkgraph.sh --mono data/lang exp/$expdir exp/$expdir/graph ||
exit 1;
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd"
exp/$expdir/graph $datadir/test exp/$expdir/decode || exit 1;
local/score.sh --cmd run.pl $datadir/test exp/$expdir/graph
exp/$expdir/decode || exit 1;
```

- The triphone GMM-HMM system will be built from the alignments generated from the monophone system. Here, we have option to vary the number of senones and Gaussians in the triphone trees. The triphone GMM-HMM system can be built by the following commands:

```
expdir=mono_mfcc
tridir=tri_mfcc
  steps/align_si.sh --boost-silence 1.25 --nj "$train_nj" --cmd "$train_cmd"
   $datadir/train data/lang exp/$expdir exp/$expdir_ali || exit 1;
for sen in 1800 2000 2200 2500; do
for gauss in 12 14 16; do
  gauss=$(($sen * $gauss))
  steps/train_deltas.sh --cmd "$train_cmd" $sen $gauss $datadir/train
  data/lang exp/$expdir_ali exp/$tridir_$sen_$gauss || exit 1;
  utils/mkgraph.sh data/lang exp/$tridir_$sen_$gauss
  exp/$tridir_$sen_$gauss/graph || exit 1;
  steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd"
  exp/$tridir_$sen_$gauss/graph $datadir/test
exp/$tridir_$sen_$gauss/decode || exit 1;
```

- The triphone system with the lowest % WER is selected for the LDA+MLLT system building. For example, here a system with 2000 senones and 12 Gaussians is selected.

```
steps/align_si.sh --nj "$train_nj" --cmd "$train_cmd" data/train data/lang
exp/tri_mfcc_2000_24000 exp/tri_mfcc_2000_24000_ali || exit 1;
for sen in 2000 2500 3000; do
for gauss in 12 16; do
  gauss=$(($sen * $gauss))
  steps/train_lda_mllt.sh --cmd "$train_cmd" --splice-opts
  "--left-context=3 --right-context=3" $sen $gauss data/train data/lang
  exp/tri_mfcc_2000_24000_ali   exp/$tridir2_$sen_$gauss || exit 1;
  utils/mkgraph.sh data/lang exp/$tridir2_$sen_$gauss
  exp/$tridir2_$sen_$gauss/graph2 || exit 1;
  steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd"
  exp/$tridir2_$sen_$gauss/graph2 data/test
exp/$tridir2_$sen_$gauss/decode3 || exit 1;
done
done
```

D.5 Acoustic Modeling using DNN-HMM

- The LDA-MLLT system with the lowest % WER is selected for the hybrid DNN-HMM experiments. First generate the alignments from the LDA-MLLT system as follows:

```
expdir=exp/tri2_mfcc_2500_40000
steps/align_si.sh --nj 8 --cmd "$train_cmd"  data/train data/lang
exp/$expdir exp/$expdir_ali
```

- To train the DNN-HMM system, the Mel filterbank features are extracted as follows:

```
fbankdir=fbank
for x in test train; do
  steps/make_fbank.sh --cmd "$train_cmd" --nj 10 $datadir/$x
  exp/makefbank/$x $fbankdir || exit 1;
  steps/compute_cmvn_stats.sh $datadir/$x exp/makefbank/$x
  $fbankdir || exit 1;
done
```

- The hybrid DNN-HMM system using nnet3 setup in the KALDI toolkit. We show a demo of building TDNN system with different numbers of hidden units as follows:

```
for x in 500 600 700 800 900; do
  datadir=nnet2_data_$features
  nndir=tri2_fbank40_TDNN_$x
  steps/nnet3/tdnn/train.sh --relu-dim $x $datadir/train
  data/lang $expdir_ali exp/tdnn/$nndir_nnet3 || exit 1;

  steps/nnet3/decode.sh $expdir/graph $datadir/test
  exp/tdnn/$nndir_nnet3/decode

  local/score.sh --cmd run.pl $datadir/test $expdir/graph
  exp/tdnn/$nndir_nnet3/decode
done
```

Bibliography

[1] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[2] Z. Wu, T. Kinnunen, N. W. D. 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, September 6-10, 2015, pp. 2037–2041.

[3] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Virginia, USA, September 8-11, 2015, pp. 1–6.

[4] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. A. L. Thomsen, A. K. Sarkar, Z. H. Tan, H. Delgado, M. Todisco *et al.*, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, March 5-9, 2017, pp. 5395–5399.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 Challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 1–6.

[6] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, June 26-29, 2018, pp. 296–303.

[7] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH*, Graz, Austria, September 15-19, 2019, pp. 1008–1012.

[8] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," in *INTERSPEECH*, Graz, Austria, September 15-19, 2019, pp. 2355–2359.

[9] M. R. Kamble and H. A. Patil, "Detection of replay spoof speech using Teager energy feature cues," *in special issue on Advances in Automatic Speaker Verification Anti-spoofing in Computer Speech and Language, Elsevier, 65 (2021): 101140.*, vol. 65, pp. 101–140, 2021.

[10] M. R. Kamble, A. K. S. Pulikonda, S. K. Maddala, and H. A. Patil, "Analysis of Teager energy profiles for spoof speech detection," *in Odyssey*, pp. 304–311, Tokyo, Japan, 2020.

[11] 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, May 12-17, 2019, pp. 2607–2611.

[12] M. R. Kamble and H. A. Patil, "Amplitude and frequency modulation-based features for detection of replayed spoof speech," *in Speech Communication, Elsevier*, vol. 125, pp. 114–127, 2020.

[13] M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," in *IEEE European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, August 28-September 2, 2017, pp. 106–110.

[14] M. R. Kamble and H. A. Patil, "Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection," in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 646–650.

[15] M. R. Kamble, A. K. S. Pulikonda, S. K. Maddala, A. Patil, R. Acharya, and H. A. Patil, "Speech demodulation-based techniques for replay and presentation attack detection," *in Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, pp. 1545–1550, Lanzhou, China, November 18-21, 2019.

[16] M. R. Kamble and H. A. Patil, "Novel variable length Teager energy profiles for replay spoof detection," *in Odyssey*, pp. 143–150, Tokyo, Japan, 2020.

[17] M. R. Kamble, S. K. Maddala, H. Tak, and H. A. Patil, "Comparison of frame and utterance-level classifers for replay attack detection," *submitted in INTERSPEECH*, Shanghai, China, October 26-29, 2020.

[18] M. R. Kamble and H. A. Patil, "Amplitude weighted frequency modulation features for spoof speech detection," *in Journal of Signal Processing Systems (JSPS)*, vol. 92, no. 8, pp. 777–791, 2020.

[19] M. R. Kamble, S. Nayak, M. A. B. Shaik, S. Rath, V. Vij, and H. A. Patil, "Teager energy spectral features for near and far-field automatic speech recognition ASR," in *submitted in IEEE European Signal Processing Conference (EUSIPCO), Dublin, Ireland*, Aug 23-27, 2021.

[20] M. R. Kamble, H. A. Patil, M. A. B. Shaik, and V. Vij, "Smoothed Teager energy features for replay spoof detection," in *submitted in IEEE European Signal Processing Conference (EUSIPCO), Dublin, Ireland*, Aug 23-27, 2021.

[21] G. P. Prajapati, , M. R. Kamble, and H. A. Patil, "Energy separation based features for replay spoof detection for voice assistant," in *in IEEE European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands*, January 1-22, 2021, pp. 386–390.

[22] K. Khoria, M. R. Kamble, and H. A. Patil, "Teager energy cepstral coefficients for classification of normal *vs.* whisper speech," in *in IEEE European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands*, January 1-22, 2021, pp. 1–5.

[23] M. R. Kamble, M. V. S. Krishna, A. K. S. Pulikonda, and H. A. Patil, "Novel Teager energy based subband features for audio acoustic scene detection and classification," in *Bhabesh Deka et. al. (Eds.) Pattern Recognition and Machine Intelligence (PReMI), Lecture Notes in Computer Science (LNCS)*, vol. 11941. Tezpur, India: Springer, December 5-8, 2019, pp. 436–444.

[24] H. Li, H. A. Patil, and M. R. Kamble, "Tutorial on spoofing attack of speaker recognition," in *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, Kuala Lumpur, Malaysia, December 12-15, 2017.

[25] T. Houtgast and H. J. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America (JASA)*, vol. 77, no. 3, pp. 1069–1077, 1985.

[26] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 118, no. 2, pp. 887–906, 2005.

[27] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 641–645.

[28] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 12–16.

[29] M. R. Kamble and H. A. Patil, "Novel amplitude weighted frequency modulation features for replay spoof detection," *in International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 185–189, Taipei, Taiwan, November 26-29, 2018.

[30] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, June 26-29, 2018, pp. 312–319.

[31] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, September 22-25, 1997, pp. 1895–1898.

[32] Y. Gong and C. Poellabauer, "Protecting voice controlled systems using sound source identification based on acoustic cues," in *IEEE 27^{th} International Conference on Computer Communication and Networks (ICCCN)*, Hangzhou, China, July 30 - August 2, 2018, pp. 1–9.

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

[34] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Voice antispoofing," *Handbook of Biometric Antispoofing, S. Marcel, SZ Li, and M. Nixon, Eds. Springer*, 2014.

[35] Y. Gong and C. Poellabauer, "An overview of vulnerabilities of voice controlled systems," *arXiv preprint arXiv:1803.09156*, 2018.

[36] H. Dai, W. Wang, A. X. Liu, K. Ling, and J. Sun, "Speech based human authentication on smartphones," in 16^{th} *Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, June 10-13, 2019, pp. 1–9.

[37] R. A. Aqeel-ur Rehman and H. Khursheed, "Voice controlled home automation system for the elderly or disabled people," *Journal of Applied Environmental and Biological Sciences*, vol. 4, pp. 55–64, 2014.

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

[39] J. Koppell, "International organization for standardization," *Handb Transnatl Gov Inst Innov*, vol. 41, p. 289, 2011.

[40] D. Markham, *Phonetic Imitation, Accent, and the Learner*. Linguistics and Phonetics, 1997, vol. 33.

[41] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, October 20-24, 2004, pp. 145–148.

[42] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry," in *INTERSPEECH*, Lyon, France, August 25-29, 2013, pp. 930–934.

[43] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, Georgia, USA, May 7-10, 1996, pp. 373–376.

[44] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[45] Y. Qian, F. K. Soong, and Z.-J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 280–290, 2013.

[46] J. F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates." in *INTERSPEECH*, Antwerp, Belgium, August 27-31, 2007, pp. 2053–2056.

[47] T. Kinnunen, Z. Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 20-25, 2012, pp. 4401–4404.

[48] J. Lindberg, M. Blomberg *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques," in *EUROSPEECH*, vol. 99, Budapest, Hungary, September 5-9, 1999, pp. 1211–1214.

[49] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA Workshop*, Vigo, Spain, November 10-12, 2010, pp. 131–134.

[50] J. Villalba and E. Lleida, “Detecting replay attacks from far-field recordings on speaker verification systems,” in *European Workshop on Biometrics and Identity Management*, Roskilde, Denmark, March 8-10, 2011, pp. 274–285.

[51] P. Maragos, T. F. Quatieri, and J. F. Kaiser, “On separating amplitude from frequency modulations using energy operators,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, San Francisco, California, USA, March 23-26, 1992, pp. 1–4.

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

[53] D. Dimitrios, M. Petros, and P. Alexandros, “Auditory Teager energy cepstrum coefficients for robust speech recognition.” in *INTERSPEECH*, Lisboa, Portugal, 2005, pp. 3013–3016.

[54] I. Rodomagoulakis and P. Maragos, “Improved frequency modulation features for multichannel distant speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 841–849, 2019.

[55] D. T. Grozdic and S. T. Jovicic, “Whispered speech recognition using deep denoising autoencoder and inverse filtering,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2313–2322, 2017.

[56] B. R. Marković, J. Galić, and M. Mijić, “Application of Teager energy operator on linear and mel scales for whispered speech recognition,” *Archives of Acoustics*, vol. 43, 2018.

[57] P. Maragos, J. F. Kaiser and T. F. Quatieri, “Energy separation in signal modulations with application to speech analysis,” *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[58] M. R. Kamble, H. Tak, S. K. Maddala, and H. A. Patil, “Novel demodulation-based features using classifier-level fusion of GMM and CNN for replay detection,” *in International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 334–338, Taipei, Taiwan, November 26-29, 2018.

[59] M. R. Kamble and H. A. Patil, “Effectiveness of mel scale-based ESA-IFCC features for classification of natural vs. spoofed speech,” *B.U. Shankar et. al. (Eds.) Pattern Recognition and Machine Intelligence (PReMI), Lecture Notes in Computer Science (LNCS)*, vol. 10597, pp. 308–316, Kolkata, India, December 17-20, 2017.

[60] H. A. Patil and M. R. Kamble, “A survey on replay attack detection for automatic speaker verification (ASV) system,” *in Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, pp. 1047–1053, Hawaii, USA, November 12-15, 2018.

[61] H. B. Sailor, M. V. S. Krishna, D. Chhabra, A. T. Patil, M. R. Kamble, and H. A. Patil, “DA-IICT/IIITV system for low resource speech recognition challenge 2018.” in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 3187–3191.

[62] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, “Presentation attack detection using long-term spectral statistics for trustworthy speaker verification,” in *IEEE*

International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, September 21-23, 2016, pp. 1–6.

[63] Z. Wu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

[64] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *ISCA Speech Synthesis Workshop (SSW), Sunnyvale, California, USA*, 2016, pp. 1–15.

[65] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[66] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 15-20, 2018, pp. 1–15.

[67] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," in *Odyssey, Les Sables d'Olonne, France*, June 26-29, 2018, pp. 240–247.

[68] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH*, Lyon, France, August 25-29, 2013, pp. 925–929.

[69] N. Evans, S. Z. Li, S. Marcel, and A. Ross, "Guest editorial: Special issue on biometric spoofing and countermeasures," *IEEE Transactions on Information forensics and security*, vol. 10, no. 4, pp. 699–702, 2015.

[70] N. Evans, S. Marcel, A. Ross, and A. B. J. Teoh, "Biometrics security and privacy protection [from the guest editors]," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 17–18, 2015.

[71] "JSTSP Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification," https://signalprocessingsociety.org/blog/jstsp-special-issue-spoofing-and-countermeasures, {Last Accessed: 2019-07-27}.

[72] "Special Issue on Speaker and Language Characterization and Recognition: Voice modeling, Conversion, Synthesis and Ethical Aspects," https://www.journals.elsevier.com/computer-speech-and-language/call-for-papers/special-issue-on-speaker-and-language-characterization, {Last Accessed: 2019-07-27}.

[73] "Special Issue on Advances in Automatic Speaker Verification Anti-spoofing," https://www.journals.elsevier.com/computer-speech-and-//language/call-for-papers/advances-in-automatic-speaker, {Last Accessed: 2019-07-27}.

[74] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[75] Y. W. Lau, D. Tran, and M. Wagner, “Testing voice mimicry with the YOHO speaker verification corpus,” in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Melbourne, VIC, Australia: Springer, 2005, pp. 15–21.

[76] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, “How vulnerable are prosodic features to professional imitators?” in *Odyssey The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008, pp. 1–4.

[77] A. K. Jain, S. Prabhakar, and S. Pankanti, “On the similarity of identical twin fingerprints,” *Pattern Recognition*, vol. 35, no. 11, pp. 2653–2663, 2002.

[78] L. Kersta and J. Colangelo, “Spectrographic speech patterns of identical twins,” *The Journal of the Acoustical Society of America (JASA)*, vol. 47, no. 1A, pp. 58–59, 1970.

[79] H. A. Patil and K. K. Parhi, “Variable length Teager energy based mel cepstral features for identification of twins,” in *International Conference on Pattern Recognition and Machine Intelligence*. Berlin Heidelberg, Germany: Springer, 2009, pp. 525–530.

[80] “HSBC reports high trust levels in biometric tech as twins spoof its voice ID system,” *Biometric Technology Today*, vol. 2017, no. 6, p. 12, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969476517301194

[81] “BBC fools HSBC voice recognition security system,” https://www.bbc.com/news/technology-39965545, {Last Accessed: 2018-10-15}.

[82] “Twins fool HSBC voice biometrics - BBC,” https://www.finextra.com/newsarticle/30594/twins-fool-hsbc-voice-biometrics--bbc, {Last Accessed: 2018-10-15}.

[83] “Contributions to biometric recognition: Matching identical twins and latent fingerprints,” Alessandra Aparecida Paulino, Ph.D. Thesis, Michigan State University, USA, Last Accessed: 2019-11-26.

[84] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, “Tacotron: Towards end-to-end speech synthesis,” *arXiv preprint arXiv:1703.10135*, Last Accessed: 2018-10-17, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[85] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, “SAS: A speaker verification spoofing database containing diverse attacks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, April 19-24, 2015, pp. 4440–4444.

[86] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, “Evaluation of speaker verification security and detection of HMM-based synthetic speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[87] Z. Wu and H. Li, “Voice conversion versus speaker verification: An overview,” *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[88] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[89] E.-K. Kim, S. Lee, and Y.-H. Oh, “Hidden Markov model based voice conversion using dynamic characteristics of speaker,” in *European Conference on Speech Communication and Technology*, Rhodes, Greece, September 22-25, 1997, pp. 1–4.

[90] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, “Text-independent voice conversion based on unit selection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 14-19, 2006, pp. I–81–I–84.

[91] M. M. Wilde and A. B. Martinez, “Probabilistic principal component analysis applied to voice conversion,” in *IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, Pacific Grove, California, November 7-10, 2004, pp. 2255–2259.

[92] S. Zhang, D. Huang, L. Xie, E. S. Chng, H. Li, and M. Dong, “Non-negative matrix factorization using stable alternating direction method of multipliers for source separation,” in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, Hong Kong, December 12-16, 2015, pp. 222–228.

[93] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 19-24, 2009, pp. 3893–3896.

[94] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[95] D. Erro and A. Moreno, “Weighted frequency warping for voice conversion,” in *INTERSPEECH*, Antwerp, Belgium, August 27-31, 2007, pp. 1965–1968.

[96] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, “ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge,” {Last Accessed = 15 Oct 2018)}. [Online]. Available: http://www.asvspoof.org/

[97] J. P. Campbell, “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[98] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, “A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case,” in *IEEE Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Hollywood, California, December 3-6, 2012, pp. 1–5.

[99] P. L. De Leon, M. Pucher, and J. Yamagishi, “Evaluation of the vulnerability of speaker verification to synthetic speech,” in *Odyssey The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 28-July 1, 2010, pp. 1–8.

[100] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. Mello, R. Violato, F. Simões, M. Uliani Neto, M. de Assis Angeloni, J. A. Stuchi *et al.*, “Overview of BTAS 2016 speaker anti-spoofing competition,” IDIAP, Tech. Rep., September 6-9, 2016.

[101] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Washington DC, USA, September 20- October 02, 2013, pp. 1–8.

[102] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *IEEE Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA)*, Chiang Mai, Thailand, December 9-12, 2014, pp. 1–5.

[103] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[104] "Objective Control for TAlker VErification (OCTAVE)," https://www.octave-project.eu/, {Last Accessed: 2018-10-15}.

[105] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The RedDots data collection for speaker recognition," in *INTERSPEECH*, Dresden, Germany, September 6-10, 2015, pp. 2996–3000.

[106] T. Kinnunen *et al.*, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[107] P. L. D. Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *INTERSPEECH*, Portland, Oregon, September 9-13, 2012, pp. 370–373.

[108] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 26-31, 2013, pp. 7234–7238.

[109] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 22-27, 2011, pp. 4844–4847.

[110] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, June 21-24, 2016, pp. 249–252.

[111] T. B. Patel and H. 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, September 6-10, 2015, pp. 2062–2066.

[112] Todisco, Massimiliano and Delgado, Héctor and Evans, Nicholas, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[113] 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*, September 6-10, 2015, pp. 2052–2056.

[114] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, September 6-10, 2015, pp. 2087–2091.

[115] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 651–655.

[116] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, "Front-end for antispoofing countermeasures in speaker verification: scattering spectral decomposition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 632–643, 2017.

[117] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection- the SJTU system for ASVspoof 2015 challenge," in *INTERSPEECH, Dresden, Germany*, 2015, pp. 2052–2056.

[118] C. Zhang, C. Yu, and J. H. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[119] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication, Elsevier*, vol. 85, pp. 43–52, 2016.

[120] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Odyssey*, Bilbao, Spain, June 21-24, 2016, pp. 270–276.

[121] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Unsupervised representation learning using convolutional restricted Boltzmann machine for spoof speech detection," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 2601–2605.

[122] H. Yu, Z.-H. Tan, Y. Zhang, Z. Ma, and J. Guo, "DNN filter bank cepstral coefficients for spoofing detection," *IEEE Access*, vol. 5, pp. 4779–4787, 2017.

[123] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-term spectral statistics for voice presentation attack detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2098–2111, Nov. 2017.

[124] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, July 2017.

[125] T. Heittola, E. Çakır, and T. Virtanen, "The machine learning approach for analysis of sound scenes and events,," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 13–40.

[126] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 5-9, 2017, pp. 2721–2725.

[127] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 15-20, 2018, pp. 1–5.

[128] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-end convolutional neural network-based voice presentation attack detection," in *IEEE International Joint Conference on Biometrics (IJCB)*, Denver, Colorado, USA, October 1-4, 2017, pp. 335–341.

[129] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform CLDNNS," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 5-9, 2017, pp. 4860–4864.

[130] H. Dinkel, Y. Qian, and K. Yu, "Investigating raw wave deep neural networks for end-to-end speaker spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1–13, 2018.

[131] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *INTERSPEECH*, Dresden, Germany, September 6-10, 2015, pp. 2067–2071.

[132] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 1–12, 2017.

[133] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Adam's Mark Hotel Dallas, TX, USA, March 14-19, 2010, pp. 1678–1681.

[134] 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, August 20-24, 2017, pp. 7–11.

[135] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 27–31.

[136] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 32–36.

[137] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 107–111.

[138] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 22–26.

[139] P. A. Tapkir and H. A. Patil, "Novel empirical mode decomposition cepstral features for replay spoof detection," in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 721–725.

[140] T. Gunendradasan, S. Irtza, E. Ambikairajah, and J. Epps, "Transmission line cochlear model based am-fm features for replay attack detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6136–6140.

[141] B. Wickramasinghe, E. Ambikairajah, J. Epps, V. Sethu, and H. Li, "Auditory inspired spatial differentiation for replay spoofing attack detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6011–6015.

[142] A. T. Patil, A. Rajul, P. A. K. Sai, and H. A. Patil, "Energy sepration-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection," *accepted in INTERSPEECH*, pp. 1–5, Graz, Austria, September 15-19, 2019.

[143] H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection," in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 726–730.

[144] K. Sriskandaraja, G. Suthokumar, V. Sethu, and E. Ambikairajah, "Investigating the use of scattering coefficients for replay attack detection," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, Malaysia, December 12-15, 2017, pp. 1195–1198.

[145] K. Srinivas and H. A. Patil, "Relative phase shift features for replay spoof detection system," in *Spoken Language Technologies for Under-resourced languages (SLTU)*, Gurugram, India, August 29-31, 2018, pp. 1–5.

[146] M. S. Saranya, R. Padmanabhan, and H. A. Murthy, "Replay attack detection in speaker verification using non-voiced segments and decision level feature switching," in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, Indian Institute of Science (IISc), Bangalore, July 16-19, 2018, pp. 1–5.

[147] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Phoneme specific modelling and scoring techniques for anti spoofing system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6106–6110.

[148] M. Liu, L. Wang, J. Dang, S. Nakagawa, H. Guan, and X. Li, "Replay attack detection using magnitude and phase information with attention-based adaptive filters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6201–6205.

[149] 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*, Stockholm, Sweden, August 20-24, 2017, pp. 17–21.

[150] 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.

[151] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 102–106.

[152] 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.

[153] H. B. Sailor, M. R. Kamble, and H. A. Patil, “Auditory filterbank learning for temporal modulation features in replay spoof speech detection,” *INTERSPEECH*, pp. 666–670, Hyderabad, India, September 2-6, 2018.

[154] G. Valenti, H. Delgado, M. Todisco, N. Evans, and L. Pilati, “An end-to-end spoofing countermeasure for automatic speaker verification using evolving recurrent neural networks,” in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d’Olonne, France, June 26-29, 2018, pp. 288–295.

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

[156] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, “Attentive filtering networks for audio replay attack detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6316–6320.

[157] P. A. Tapkir, M. R. Kamble, and H. A. Patil, “Replay spoof detection using power function based features,” in *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, Hawaii, USA, November 12-15, 2018, pp. 1019–1023.

[158] M. J. Alam, G. Bhattacharya, and P. Kenny, “Boosting the performance of spoofing detection systems on replay attacks using q-logarithm domain feature normalization,” in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d’Olonne, France, June 26-29, 2018, pp. 393–398.

[159] A. Janicki, “Spoofing countermeasure based on analysis of linear prediction error,” in *INTERSPEECH*, Dresden, Germany, September 6-10, 2015.

[160] Y. Stylianou, “Removing linear phase mismatches in concatenative speech synthesis,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 232–239, 2001.

[161] W. Kim and J. H. L. Hansen, “Angry emotion detection from real-life conversational speech by leveraging content structure,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5166–5169, March 14-19, 2010.

[162] H. A. Patil and T. Basu, “Teager energy Mel cepstrum for identification of twins in Marathi,” in *IEEE INDICON*, Kharagpur, India, December 20-22, 2004, pp. 58–61.

[163] C. Jankowski, T. F. Quatieri, and D. A. Reynolds, “Measuring fine structure in speech: Application to speaker identification,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Detroit, Michigan, USA, May 08-12, 1995, pp. 325–328.

[164] F. Jabloun, A. E. Cetin, and E. Erzin, “Teager energy based feature parameters for speech recognition in car noise,” *IEEE Signal Processing Letters*, vol. 6, no. 10, pp. 259–261, 1999.

[165] J. F. Kaiser, “On a simple algorithm to calculate the energy of a signal,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, USA, April 3-6, 1990, pp. 381–384.

[166] K. Vijayan, V. Kumar, and K. S. R. Murty, "Feature extraction from analytic phase of speech signals for speaker verification." in *INTERSPEECH*, Singapore, September 14-18, 2014, pp. 1658–1662.

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

[168] H. A. Patil and K. K. Parhi, "Development of TEO phase for speaker recognition," in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, Indian Institute of Science (IISc), Bangalore, July 18-21, 2010, pp. 1–5.

[169] L. Deng and D. O'Shaughnessy, *Speech Processing – A Dynamic and Optimization-Oriented Approach*. 1^{st} Edition, Marcel Dekker Inc., June 2003.

[170] S. Mallat, *A Wavelet Tour of Signal Processing. 2^{nd} Edition*. Academic press, 1999.

[171] J. Klapper and C. Harris, "On the response and approximation of Gaussian filters," *IEEE IRE Transactions on Audio*, vol. 7, no. 3, pp. 80–87, 1959.

[172] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[173] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America (JASA)*, vol. 8, no. 3, pp. 185–190, 1937.

[174] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 1, Hong Kong, China, April 6-10, 2003, pp. I–656–659–I.

[175] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice. 1^{st} Edition*. Pearson Education India, 2006.

[176] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in ASV systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, November 14-16, 2017, pp. 51–55.

[177] "Reverberation," https://byjus.com/physics/reverberation/, {Last Accessed: 2018-10-24}.

[178] B. Blesser and L.-R. Salter, *Spaces Speak, Are You Listening?: Experiencing Aural Architecture*. MIT Press, 2009.

[179] H. Kuttruff, *Room Acoustics*. 1^{st} Edition, CRC Press, 2016.

[180] R. Kuc, *Introduction to Digital Signal Processing*. 1^{st} Edition, McGraw-Hill, Inc., 1988.

[181] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1–19, 2016.

[182] B. Boashash, *Time-Frequency Signal Analysis. 2^{nd} Edition*. Prentice Hall, 1991.

[183] L. Cohen, *Time-Frequency Analysis. 1^{st} Editon*. Prentice Hall PTR Englewood Cliffs, NJ:, 1995, vol. 778.

[184] I. Arroabarren, X. Rodet, and A. Carlosena, “On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1413–1421, 2006.

[185] J. Traer and J. H. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[186] H. Tak and H. Patil, “Novel linear frequency residual cepstral features for replay attack detection,” in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 726–730.

[187] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, and M. Todisco, “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.

[188] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, “Short-time instantaneous frequency and bandwidth features for speech recognition,” in *IEEE Signal Processing Society workshop on Automatic Speech Recognition and Understanding (ASRU)*, Merano Kurhaus Merano, Italy, December 13-17, 2009, pp. 103–106.

[189] S. C. Gupta, “Phase-locked loops,” *Proceedings of the IEEE*, vol. 63, no. 2, pp. 291–306, 1975.

[190] 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.

[191] K. S. R. Murty and B. Yegnanarayana, “Epoch extraction from speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[192] M. Alam, P. Ouellet, P. Kenny, and D. O’Shaughnessy, “Comparative evaluation of feature normalization techniques for speaker verification,” *Advances in Nonlinear Speech Processing*, pp. 246–253, 2011.

[193] A. A. Garcia and R. J. Mammone, “Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 1, Phoenix, Arizona, USA, March 15-19, 1999, pp. 325–328.

[194] M. Westphal, “The use of cepstral means in conversational speech recognition,” in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, September 22-25, 1997, pp. 1143–1146.

[195] J. Schnupp, I. Nelken, and A. King, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, First Edition, 2012.

[196] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, “Linear versus mel frequency cepstral coefficients for speaker recognition,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, December 11-15, 2011, pp. 559–564.

[197] A. Potamianos and P. Maragos, “A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation,” *Signal processing*, vol. 37, no. 1, pp. 95–120, 1994.

[198] R. Kaszynski and J. Piskorowski, “New concept of delay equalized low-pass Butterworth filters,” in *IEEE International Symposium on Industrial Electronics*, vol. 1, Montreal, Quebec, Canada, July -13, 2006, pp. 171–175.

[199] S. Jelil, S. Kalita, S. M. Prasanna, and R. Sinha, “Exploration of compressed ILPR features for replay attack detection,” in *INTERSPEECH*, Hyderabad, India, September 2-6, 2018, pp. 631–635.

[200] J. Yang, R. K. Das, and H. Li, “Extended constant-Q cepstral coefficients for detection of spoofing attacks,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), November 12-15, 2018*. Hawaii, USA: IEEE, pp. 1024–1029.

[201] R. K. Das and H. Li, “Instantaneous phase and excitation source features for detection of replay attacks,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), November 12-15, 2018*. Hawaii, USA: IEEE, pp. 1030–1037.

[202] P. B. Bachhav, H. A. Patil, and T. B. Patel, “A novel filtering based approach for epoch extraction,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, April 19-24, 2015, pp. 4784–4788.

[203] V. Tomar and H. A. Patil, “On the development of variable length Teager energy operator (VTEO),” in *INTERSPEECH*, Brisbane, Queensland, Australia, pp. 1056–1059.

[204] H. A. Patil and K. K. Parhi, “Novel variable length Teager energy based features for person recognition from their hum,” in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Adam’s Mark Hotel Dallas, TX, USA, March 14-19, 2010, pp. 4526–4529.

[205] J. Choi and T. Kim, “Neural action potential detector using multi-resolution TEO,” *Electronics Letters*, vol. 38, no. 12, pp. 541–543, 2002.

[206] A. Potamianos and P. Maragos, “Speech formant frequency and bandwidth tracking using multiband energy demodulation,” *The Journal of the Acoustical Society of America (JASA)*, vol. 99, no. 6, pp. 3795–3806, 1996.

[207] D. Dimitriadis, P. Maragos, and A. Potamianos, “Auditory Teager energy cepstrum coefficients for robust speech recognition,” in *European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Portugal, September 4-8, 2005, pp. 3013–3016.

[208] F. Jabloun, A. E. Cetin, and E. Erzin, “Teager energy based feature parameters for speech recognition in car noise,” *IEEE Signal Processing Letters*, vol. 6, no. 10, pp. 259–261, 1999.

[209] R. Sharma, L. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner, and S. Prasanna, “Empirical mode decomposition for adaptive AM-FM analysis of speech: A review,” *Speech Communication*, vol. 88, pp. 39–64, 2017.

[210] D. Dimitriadis and E. Bocchieri, "Use of micro-modulation features in large vocabulary continuous speech recognition tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1348–1357, 2015.

[211] 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.

[212] H. Luo, Y. Wang, D. Poeppel, and J. Z. Simon, "Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence," *Journal of Neurophysiology*, vol. 96, no. 5, pp. 2712–2723, 2006.

[213] L. Cohen, K. Assaleh, and A. Fineberg, "Instantaneous bandwidth and formant bandwidth," in *IEEE SP Workshop on Statistical Signal and Array Processing*, Quebec, City, QC, Canada, October 7-9, 1992, pp. 13–17.

[214] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press, 2014.

[215] B. Chettri and B. L. Sturm, "A deeper look at Gaussian mixture model based anti-spoofing systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 15-20, 2018, pp. 5159–5163.

[216] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[217] A. Martin *et al.*, "The DET curve in assessment of decision task performance," in *European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, September 22-25, 1997, pp. 1895–1898.

[218] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Orlando, Florida, USA, May 13-17, 2002, pp. IV–4072.

[219] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[220] S. Dupont, C. Ris, and D. Bachelart, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, UK, 2004.

[221] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 504–511.

[222] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted boltzmann machine and Teager energy operator for speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 141, no. 6, pp. EL500–EL506, 2017.

[223] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia, April 19-24, 2015, pp. 5206–5210.

[224] D. Povey *et al.*, “The KALDI speech recognition toolkit,” in *IEEE Signal Processing Society workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, USA, Drecember 11-15, 2011, pp. 1–5.

[225] H. Xu, D. Povey, L. Mangu, and J. Zhu, “Minimum Bayes risk decoding and system combination based on a recursion for edit distance,” *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[226] P. Agrawal and S. Ganapathy, “Modulation filter learning using deep variational networks for robust speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 244–253, 2019.

[227] N. Ma, R. Marxer, J. Barker, and G. J. Brown, “Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition,” in *IEEE Signal Processing Society workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 490–495.

[228] L. Pfeifenberger, T. Schrank, M. Zohrer, M. Hagmüller, and F. Pernkopf, “Multichannel speech processing architectures for noise robust speech recognition: 3rd chime challenge results,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 452–459.

[229] D. Bagchi *et al.*, “Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition,” in *IEEE Signal Processing Society workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 496–503.

[230] T. Hori *et al.*, “The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 475–481.

[231] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, “The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 436–443.

[232] S. Sivasankaran *et al.*, “Robust ASR using neural network based speech enhancement and feature simulation,” in *IEEE Signal Processing Society workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 13-17, 2015, pp. 482–489.

[233] L. Blue, L. Vargas, and P. Traynor, “Hello, is it me you’re looking for?: Differentiating between human and electronic speakers for voice interface security,” in 11th *ACM Conference on Security & Privacy in Wireless and Mobile Networks*, June 18-20, 2018, pp. 123–133.

[234] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Smoothed nonlinear energy operator-based amplitude modulation features for robust speech recognition," in *International Conference on Nonlinear Speech Processing*. Mons, Belgium: Springer, June 19-21, 2013, pp. 168–175.

[235] H. Beyramienanlou and N. Lotfivand, "An efficient Teager energy operator-based automated QRS complex detection," *Journal of Healthcare Engineering*, 2018.

[236] R. W. Morris, "Enhancement and recognition of whispered speech," Ph.D. dissertation, Georgia Institute of Technology, 2003.

[237] C.-Y. Yang, G. Brown, L. Lu, J. Yamagishi, and S. King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with vts compensation," in *IEEE 8th International Symposium on Chinese Spoken Language Processing*, Hong Kong, China, 2012, pp. 220–223.

[238] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[239] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[240] A. Mathur, S. M. Reddy, and R. M. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 157, 2012.

[241] S. Ghaffarzadegan, H. Boşil, and J. H. Hansen, "Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, April 19-24, 2015, pp. 5024–5028.

[242] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *International Conference on Spoken Language Processing (ISCSLP)*, Jeju Island, Korea, pp. 1493–1496.

[243] C. Zhang and J. H. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 883–894, 2010.

[244] S. Ghaffarzadegan, H. Bořil, and J. H. Hansen, "Ut-vocal effort ii: Analysis and constrained-lexicon recognition of whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 2544–2548.

[245] P. X. Lee, D. Wee, H. S. Y. Toh, B. P. Lim, N. F. Chen, and B. Ma, "A whispered mandarin corpus for speech technology applications," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[246] S. T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acustica united with Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[247] J. B. Wilson and J. D. Mosko, "A comparative analysis of whispered and normally phonated speech using an LPC-10 vocoder," Rome Air Development Center Giffis AFB NY, Tech. Rep., 1985.

[248] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech, Language and the Law*, vol. 12, no. 2, pp. 143–173, 2005.

[249] C. Huang, X. Y. Tao, L. Tao, J. Zhou, and H. B. Wang, "Reconstruction of whisper in chinese by modified melp," in *7th International Conference on Computer Science & Education (ICCSE)*, Melbourne, Australia., July 4-6, 2012, pp. 349–353.

[250] I. V. Mcloughlin, H. R. Sharifzadeh, S. L. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, pp. 1–21, 2015.

[251] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 515–520, 2002.

[252] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[253] V.-A. Tran, G. Bailly, H. Loevenbruck, and T. Toda, "Improvement to a nam-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314–326, 2010.

[254] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[255] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba, and A. Moinet, "Voice conversion for whispered speech synthesis," *IEEE Signal Processing Letters*, vol. 27, no. 01, pp. 186–190, 2019.

[256] J. H. Hansen, C. Zhang, and X. Fan, "Speech processing for robust speaker recognition: Analysis and advancements for whispered speech," in *Forensic Speaker Recognition, Neustein and Patil (Eds.)*. Springer, 2011, pp. 253–272.

[257] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[258] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[259] G. Zhou, J. Hansen, and J. Kaiser, "Classification of speech under stress based on features derived from the nonlinear teager energy operator," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, USA, May 12-15, 1998, pp. 549—-552.

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

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

[262] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[263] Bugalho, Miguel and Portelo, José and Trancoso, Isabel and Pellegrini, Thomas and Abad, Alberto, "Detecting audio events for semantic video search," in *INTERSPEECH*, Brighton, United Kingdom, (UK), September 6-9, 2009, pp. 1151–1154.

[264] Eronen, Antti J and Peltonen, Vesa T and Tuomi, Juha T and Klapuri, Anssi P and Fagerlund, Seppo and Sorsa, Timo and Lorho, Gaëtan and Huopaniemi, Jyri, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing,*, pp. 321–329, 2006.

[265] Valenti, Michele and Diment, Aleksandr and Parascandolo, Giambattista and Squartini, Stefano and Virtanen, Tuomas, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Workshop Detection Classification Acoustic Scenes Events*, Budapest, Hungary, November 2-3, 2016, pp. 95–99.

[266] Valenti, Michele and Squartini, Stefano and Diment, Aleksandr and Parascandolo, Giambattista and Virtanen, Tuomas, "A convolutional neural network approach for acoustic scene classification," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, July 14-19, 2017, pp. 1547–1554.

[267] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, Lille, France, July 6-11, 2015, pp. 448–456.

[268] H. A. Patil, P. Dutta, and T. Basu, "Effectiveness of LP based features for identification of professional mimics in indian languages," in *Int Workshop on Multimodal User Authentication, MMUA06, Toulouse, France*, 2006, pp. 11–18.

[269] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, Second Edition, 2004.

[270] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters, Elsevier*, vol. 27, no. 8, pp. 861–874, 2006.

[271] W. J. Youden, "Index for rating diagnostic tests," *Cancer, Wiley Subscription Services, Inc., A Wiley Company*, vol. 3, no. 1, pp. 32–35, 1950.

[272] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA), Protein Structure, Elsevier*, vol. 405, no. 2, pp. 442–451, 1975.

[273] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., First Edition, 1993.

[274] H. A. Patil and S. Viswanath, "Effectiveness of Teager energy operator for epoch detection from speech signals," *International Journal Speech Technology (IJST), Springer*, vol. 14, no. 4, pp. 321–337, Dec. 2011.

[275] H. A. Patil and M. C. Madhavi, "Combining evidences from magnitude and phase information using VTEO for person recognition using humming," *Computer Speech and Language, Elsevier*, vol. 52, pp. 225–256, 2018.

List of Publications from Thesis

International Journal Papers

1. **Madhu R. Kamble**, Hardik. B. Sailor, Hemant. A. Patil and Haizhou. Li, “Advances in Anti-spoofing: From the Perspective of ASVspoof Challenges," in APSIPA Transactions on Signal and Information Processing vol. 9, 2020. **(Invited Paper)**.
2. **Madhu R. Kamble** and Hemant. A. Patil, “Amplitude Weighted Frequency Modulation Features for Spoof Speech Detection,” in Journal of Signal Processing Systems (JSPS) 92 (8), 777-791, 2020 **(Invited Paper)**.
3. **Madhu R. Kamble** and Hemant. A. Patil, “Detection of Replay Spoof Speech Using Teager Energy Feature Cues Detection," in special issue on Advances in Automatic Speaker Verification Anti-spoofing in Computer Speech and Language, Elsevier, 65 (2021): 101140.
4. **Madhu R. Kamble**, T. Hemlata, and Hemant A. Patil, “Amplitude and Frequency Modulation-Based Features for Detection of Replayed Spoof Speech,” in Speech Communication, Elsevier, 125 (2020): 114-127.

Book Chapters in Coedited Book Volumes

1. **Madhu R. Kamble** and Hemant. A. Patil, "Effectiveness of Mel Scale-Based ESA-IFCC Features for Classification of Natural *vs.* Spoofed Speech " in 7th in B.U. Shankar et. al. (Eds.) International Conference on Pattern Recognition and Machine Intelligence (**PReMI**), Lecture Notes in Computer Science (LNCS). Springer, vol. 10597, pp. 308–316, Kolkata, India, December 17-20, 2017.

2. **Madhu R. Kamble**, Maddala Venkata Siva Krishna, Aditya Krishna Sai Pulikonda and Hemant A. Patil, "Novel Teager Energy Based Subband Features for Audio Acoustic Scene Detection and Classification", in 8th in Bhabesh Deka et. al. (Eds.) International Conference on Pattern Recognition and Machine Intelligence (**PReMI**), Lecture Notes in Computer Science (LNCS). Springer, vol. 11941, pp. 436-444, Tezpur University (TU), Tezpur, India, December 5-8, 2019.

Conference Papers

1. **Madhu R. Kamble**, Hemant A. Patil, M. Ali Basha Shaik, and Vikram Vij, “Smoothed Teager Energy features for Replay Spoof Detection" submitted for possible publications in European Signal Processing Conference **EUSIPCO**, Dublin, Ireland 2021.

2. Dipesh K. Singh, Divyesh G. Rajpura, **Madhu R. Kamble**, Hemant A. Patil, “Smooth Filtered Instantaneous Amplitude Features for Far-Field Speaker Verification” submitted for possible publications in European Signal Processing Conference **EUSIPCO**, Dublin, Ireland 2021.

3. **Madhu R. Kamble** and Hemant A. Patil, “The Impact of Room Acoustics on Replay Speech Signal”, submitted for possible publications in European Signal Processing Conference **EUSIPCO**, Dublin, Ireland 2021.

4. **Madhu R. Kamble**, Shekhar Nayak, M. Ali Basha Shaik, Shakti Rath, Vikram Vij, and Hemant A. Patil, “Teager Energy Spectral Features for Near and Far-Field Automatic Speech Recognition (ASR)", submitted in European Signal Processing Conference **EUSIPCO**, Dublin, Ireland 2021.

5. Priyanka Gupta, Gauri P. Prajapati, Shrishti Singh, **Madhu R. Kamble**, and Hemant A. Patil, "Design of Voice Privacy System using Linear Prediction" in APSIPA-ASC, Auckland, New Zealand 2020, pp. 543-549.

6. Kuldeep Khoria, **Madhu R. Kamble**, and Hemant A. Patil, “Teager Energy Cepstral Coefficients for Classification of Normal vs. Whisper Speech", in European Signal Processing Conference (**EUSIPCO**), Amsterdam, The Netherlands pp. 1-5, 2020.

7. Gauri P. Prajapati, **Madhu R. Kamble**, and Hemant A. Patil, “Energy Separation Based Features for Replay Spoof Detection for Voice Assistant", in European Signal Processing Conference (**EUSIPCO**), Amsterdam, The Netherlands pp. 386-390, 2020.

8. **Madhu R. Kamble** and Hemant A. Patil, "Novel Variable Length Teager Energy Profiles for Replay Spoof Detection", in **Odyssey**, Tokyo, Japan, May 18-21, 2020, pp. 143-150.

9. **Madhu R. Kamble**, Aditya Krishna Sai Pulikonda, Maddala Venkata Siva Krishna, and Hemant A. Patil, "Analysis of Teager Energy Profiles for Spoof Speech Detection", in **Odyssey**, Tokyo, Japan, May 18-21, 2020, pp. 304-311.

10. **Madhu R. Kamble**, Aditya Krishna Sai Pulikonda, Maddala Venkata Siva Krishna, Ankur Patil, Rajul Acharya, and Hemant A. Patil, "Speech Demodulation-based Techniques for Replay and Presentation Attack Detection", in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA-ASC)**, Lanzhou, China, pp. 1545-1550, November 18-21, 2019.

11. **Madhu R. Kamble**, Maddala Venkata Siva Krishna, Hemlata Tak, and Hemant A. Patil, "Comparison of Frame and Utterance-level Classifiers for Replay Attack Detection", accepted in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA-ASC)**, Lanzhou, China, November 18-21, 2019.

12. **Madhu R. Kamble** and Hemant 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, pp. 2607-2611, 2019..

13. **Madhu R. Kamble**, and Hemant A. Patil, "Novel Amplitude Weighted Frequency Modulation Features for Replay Spoof Detection" in 11^{th} International Symposium on Chinese Spoken Language Processing **(ISCSLP)**, Taipei, Taiwan, pp. 185-189, November 26-29, 2018.

14. **Madhu R. Kamble**, Hemlata. Tak, Maddala Venkata Siva Krishna, and Hemant. A. Patil, "Novel Demodulation-Based Features using Classifier-level Fusion of GMM and CNN for Replay Detection" in 11^{th} International Symposium on Chinese Spoken Language Processing **(ISCSLP)**, Taipei, Taiwan, pp. 334-338, November 26-29, 2018.

15. Hemant A. Patil, **Madhu R. Kamble**, "A Survey on Replay Attack Detection for Automatic Speaker Verification (ASV) System," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA-ASC)**, Honolulu, Hawaii, USA, pp. 1047-1053, 12-15 November 2018, pp. 1047-1053.

16. Prasad Tapkir, **Madhu R. Kamble**, Hemant A. Patil, "Replay Spoof Detection using Power Function Based Features, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA-ASC)**, Honolulu, Hawaii, USA, pp. 1019-1023, 12-15 November 2018, pp. 1019-1023.

17. **Madhu R. Kamble**, “Energy Separation Algorithm based Features for Replay Spoof Detection,” in **INTERSPEECH** 2018: 4th Doctoral Consortium, IIIT-Hyderabad, September 1st 2018.

18. **Madhu R. Kamble**and Hemant A. Patil, “Novel Variable Length Energy Separation Algorithm using Instantaneous Amplitude Features For Replay Detection,” in **INTERSPEECH**, Hyderabad, India, pp. 646-650, 2018.

19. **Madhu R. Kamble**, T. Hemlata, and Hemant. A. Patil, “Effectiveness of speech demodulation-based features for replay spoof speech detection,” in **INTERSPEECH**, Hyderabad, India, pp. 641-645, 2018.

20. Hardik. B. Sailor, **Madhu R. Kamble**, and Hemant. A. Patil, “Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection,” in **INTERSPEECH**, Hyderabad, India, pp. 666-670, 2018.

21. Hardik. B. Sailor, Maddala Venkata Siva Krishna, D. Chhabra, Ankur. Patil, **Madhu R. Kamble**, and Hemant. A. Patil, “DA-IICT/IIITV System for Low Resource Speech Recognition Challenge 2018”, in **INTERSPEECH**, Hyderabad, India, pp. 3187-3191, 2018.

22. **Madhu R. Kamble** and Hemant A. Patil, “Novel Energy Separation Based Instantaneous Frequency Features for Spoof Speech Detection,” European Signal Processing Conference (**EUSIPCO**), Kos Island, Greece, Europe, pp. 116-120, 2017.

23. Hemant A. Patil, **Madhu R. Kamble**, Tanvina B. Patel and Meet Soni, “Novel Variable Length Teager Energy Separation Based IF Features for Replay Detection,” in **INTERSPEECH**, Stockholm, Sweden, pp. 12-16, 2017.

24. Hardik B. Sailor, **Madhu R. Kamble** and Hemant A. Patil, “Unsupervised Representation Learning Using Convolutional Restricted Boltzmann Machine for Spoof Speech Detection,” in **INTERSPEECH**, Stockholm, Sweden, pp. 2601-2605, 2017.

25. **Madhu R. Kamble** and Hemant. A. Patil, “Novel Energy Separation Based Frequency Modulation Features For Spoofed Speech Classification” in 9th International Conference on Advances in Pattern Recognition (**ICAPR**) , Indian Statistical Institute, Bangalore, India, 2017.

Brief Biography

Madhu R. Kamble received B.E. degree from P.V.P.I.T, Budhgaon, Sangli, Maharashtra state in 2012. She did her M.E. (with Signal Processing Specialization) in 2015 from Cummins College of Engineering, Pune, Maharashtra State, India. She was a doctoral student during July 2015-Feb 2021 at DA-IICT and currently, she is a post doctoral fellow at EURECOM, France. She has been awarded with Rajiv Gandhi National Fellowship (RGNF) from University Grants Commission (UGC) for her doctoral research studies during July 2015-April 2020.

She has published 31 research papers in top conferences and peer-reviewed journals. Her research area includes voice biometrics, Spoof Speech Detection (SSD), Voice Assistants (VAs), Automatic Speech Recognition (ASR), Whisper Speech Detection (WSD), and Acoustic Scene Classification (ASC). Her main research is focused on developing signal processing-based countermeasures and analysis of natural *vs.* spoof signals.

She offered a tutorial jointly with Prof. Patil on the Spoofing Attacks for Automatic Speaker Verification in IEEE-WIE Conference, at AISSM's Pune in Dec 2016. She was co-instructor with Prof. Patil and Prof. Haizhou Li (IEEE Fellow, ISCA Fellow) for a tutorial in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Kuala Lumpur, Malaysia, 2017. She was a research intern at Samsung Research Institute, Bangalore (SRI-B), India during May-Nov. 2019. She is a student member of ISCA, student member of IEEE, IEEE Signal Processing Society, IEEE Young Professionals, IEEE WIE, and APSIPA-ASC. She is a reviewer for Computer, Speech and Language and Nerocomputing Journal, Elsevier, IEEE Transaction on Automation Science and Engineering (ASE), and Expert Systems, Elsevier. She received ISCA and IEEE SPS student travel grant of 650 Euros and 1000 USD to present her papers during

INTERSPEECH 2017, Stockholm, Sweden, and ICASSP 2019, Brighton, UK, respectively. She was selected as student volunteer during ICASSP 2019, Brighton, UK, and awarded with 300 USD. She has been a student volunteer for ISCA supported Summer Schools, S4P 2019, S4P 2018, S4P 2017, S4P 2016, and ASAP 2016.