

Handcrafted Features for Anti-Spoofing

by

Ankur T. Patil
201621008

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



January 2023

Declaration

I hereby declare that

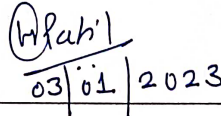
- i) the thesis comprises of my original work towards the degree of Doctor of Philosophy at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) and has not been submitted elsewhere for a degree,
- ii) due acknowledgment has been made in the text to all the reference material used.


03/01/2023

Mr. Ankur T. Patil
(Student ID : 201621008)

Certificate

This is to certify that the thesis work entitled, "*Handcrafted Features for Anti-Spoofing*," has been carried out by *Mr. Ankur T. Patil* for the degree of *Doctor of Philosophy* at *Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)* under my supervision.


03/01/2023

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

Acknowledgments

There are several living and non-living entities are responsible for certain accomplishment to happen. I take this opportunity to extend my sincere gratitude to all those who helped me directly and indirectly for making this Ph.D. thesis possible. First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisor, Prof. (Dr.) Hemant A. Patil, for his dedicated help, encouragement, and continuous support throughout my Ph.D. His passion and dedication for publishing high quality research work in top conferences and peer reviewed journals, has made a deep impression on me. During the last six years, I have learned extensively from him including how to approach a research problem by systematic thinking, how to raise a new research possibility, how to face failures, and never losing hope. I owe lots of gratitude to him for giving me his time. He always gave utmost priority to my manuscripts and Ph.D. thesis. Furthermore, I was astonished to see his deep dedication towards the academics through certain events. In particular, he ensured the conduction of the complete (even more) set of lectures and progress review meetings with lab members just after his discharge from the ICU of hospital (and had a close shave with death) during COVID-19 recovery. Also, he used to directly come to the institute from the abroad for the conduction of lectures regardless of the fatigue (especially when he suffered from acute back pain) due to long international travel. Such events will continuously inspire me to build the integrity towards the duties, and I am extremely glad to be associated with a person like him in my life.

I am also thankful to Research Progress Seminar (RPS) committee members, namely, Prof. (Dr.) K. S. Dasgupta and Prof. (Dr.) M. V. Joshi for patiently listening to my research progress at the end of each semester and providing me a valuable feedback. I am also thankful to my Ph.D. synopsis committee members, Prof. (Dr.) M. V. Joshi, Prof. (Dr.) Rajib L. Das, and Prof. (Dr.) Aditya Tatu for their valuable feedback that has helped me to revise my doctoral thesis. In addition, we thank Prof. (Dr.) Rodrigo C. Guido (Instituto de Biociências, Letras e Ciências Exatas, Unesp - Univ Estadual Paulista (São Paulo State University)) for his guidance and kind help in my two coauthored journal papers with him. I would like

to extend my thanks to then Sr. Vice President and Voice Intelligence R&D Team Head at Samsung R&D Institute, Bangalore, Dr. Vikram Vij for providing me internship opportunity with his team. It was a nice learning experience for me. I gratefully acknowledge the authorities of DA-IICT, Gandhinagar for their kind help to carry out my research work. I am also thankful to the Resource Center (RC) staff of DA-IICT for their prompt co-operations.

My heartfelt thanks to all my co-authors, Prof. (Dr.) Hemant A. Patil, Prof. (Dr.) Rodrigo C. Guido, Rajul Acharya, Kuldeep Khoria, Prasad Tapkir, Harsh Kotta, Dr. Hardik Sailor, Neil Shah, Mirali Purohit, Mihir Parmar, Sidhant Gupta, Dr. Nirmesh Shah, Dr. Madhu Kamble, Savan Doshi, Maitreya Patel, Harshit Malaviya, Aastha Kachhi, Aditya Sai, Siva Krishna Maddala, Priyanka Gupta, Piyush Chodingala, Anand Therattil, Shreya Chaturvedi, Diksha Chhabra, and Mehak Piplani for their help, support, and technical discussions for doing an excellent teamwork. My special regards to all my seniors and Ph.D. colleagues, Dr. Maulik C. Madhavi, Dr. Hardik B. Sailor, Dr. Nirmesh J. Shah, Dr. Madhu R. Kamble, and Ms. Priyanka Gupta for their guidance and moral support. I also acknowledge Divyesh, Srinivas, Sreeraj, Kirtana, Dipesh, Gauri, Shrishti for technical discussion and sharing their insights regarding research issues.

I feel a deep sense of gratitude for my aunt (Ms. Mirabai B. Patil), father (Mr. Tukaram B. Patil), and mother (Mrs. Bharati T. Patil), who formed part of my vision. Their infallible love and blessings have always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I am also very much grateful to my brother Abhijeet for motivating me for this journey and his help and technical discussions on various issues during this journey. I am very much thankful to my wife Payal for her excellent support provided during this journey. I am also thankful to my family member Jyoti and my parents-in-law Mr. Sanjay C. Patil and Ms. Anita S. Patil for always being there for help and support.

I am also thankful to the teachers of my secondary school, Kuber High School, Mhasawad (M.S., India), for their teachings and lessons. A special mention of thanks to my friends Mohan, Prashant, Kalpesh, Mahesh, Rushi, Anup, Kapil, Lalit, Pankaj, and all other friends, who contributed in mental support during my Ph.D. admission. I believe that for any desirable situation to happen, there are infinite factors (living and non-living) are responsible. Hence, I would be grateful for all those factors and one who aligned them for this wonderful journey to happen.

Ankur T. Patil

Contents

Abstract	xiii
List of Acronyms	xiii
List of Symbols	xx
List of Tables	xxii
List of Figures	xxviii
1 Introduction	1
1.1 Motivation	1
1.2 Unnaturalness in the Spoof Speech Signals	5
1.3 Development of Standard Datasets for Voice Anti-Spoofing	7
1.4 Contributions of the Thesis	8
1.4.1 Subband Filtering-Based Features	9
1.4.2 Features Derived from Spectral Representations	9
1.5 Organization of the Thesis	11
1.6 Chapter Summary	12
2 Literature Search	13
2.1 Introduction	13
2.2 Studies on SSD for SS and VC Attacks	13
2.3 Studies on Real Replay SSD	17
2.4 Studies on SSD for SS, VC, and Simulated Replay	18
2.5 Studies on SSD for SS, VC, Replay, and DeepFake Attacks	20
2.6 Other Datasets Used	21
2.7 Gap Area in the Anti-Spoofing Literature	22
2.8 Key Contributions in the Thesis	23
2.9 Chapter Summary	24
3 Experimental Setup	25
3.1 Introduction	25
3.2 Standard Anti-Spoofing Corpora	25

3.2.1	ASVSpooof 2015 Challenge Dataset	26
3.2.2	ASVSpooof 2017 Challenge Dataset	28
3.2.3	ASVSpooof 2019 Challenge Dataset	29
3.2.3.1	Logical Access (LA)	31
3.2.3.2	Physical Access (PA)	32
3.2.4	ASVSpooof 2021 Challenge Dataset	32
3.2.5	Biometrics: Theory, Applications, and Systems 2016 (BTAS 2016) Dataset	33
3.2.6	POp noise COrpus (POCO)	33
3.2.7	Realistic Replay Attack Microphone Array Speech Corpus (ReMASC)	36
3.3	Existing Feature Sets	37
3.3.1	CQCC	38
3.3.2	Cepstral Feature Sets	38
3.3.3	MFCC and LFCC	39
3.3.4	TECC and SECC	39
3.4	Classifiers Used	39
3.4.1	Gaussian Mixture Model (GMM)	39
3.4.2	Support Vector Machine (SVM)	40
3.4.3	Convolutional Neural Network (CNN)	41
3.4.4	Light-CNN	42
3.4.5	Residual Neural Networks (ResNet)	42
3.5	Performance Evaluation Metrics	43
3.5.1	Equal Error Rate (EER)	43
3.5.2	tandem - Detection Cost Function (t-DCF)	44
3.5.3	% Classification Accuracy	45
3.5.4	Area Under the Curve (AUC) for Overlapping Region	45
3.6	Score-Level Data Fusion	46
3.7	Chapter Summary	47
4	Features using TEO	49
4.1	Introduction	49
4.2	Motivation for TEO	50
4.3	Derivation of TEO and ESA	52
4.4	ETECC Feature Set	57
4.4.1	Signal Mass (ρ) and ETEO	57
4.4.2	ETECC Feature Extraction	59
4.4.3	Spectrographic Analysis	63

4.4.3.1	TEO <i>vs.</i> ETEO Profiles	63
4.4.3.2	Waterfall Plot of TECC <i>vs.</i> ETECC Feature Sets . .	64
4.4.4	Experimental Setup	64
4.4.5	Experimental Results	66
4.4.5.1	Paraconsistent Feature Engineering (PFE) for Rank- ing	66
4.4.5.2	Results on ASVSpooF 2017 Version-2 Dataset . . .	70
4.4.5.3	Results on ASVSpooF 2019 Challenge, BTAS, and AVSspooF 2015 Challenge Datasets.	81
4.4.5.4	Results on ReMASC Dataset	82
4.5	CTECC _{max} Feature Set	85
4.5.1	Cross-Teager Energy Operator (CTEO)	86
4.5.2	CTECC _{max} Feature Extraction Procedure	90
4.5.3	Experimental Setup	92
4.5.4	Spectrographic Analysis	92
4.5.5	Results on Individual Systems and Their Fusions	93
4.5.6	Results Obtained on Environment-Dependent and Environment- Independent Scenarios	95
4.5.7	Results using Individual Recording Devices	96
4.5.8	Detection Error Trade-off (DET) Curves	99
4.6	CFCCIF-ESA Feature Set	101
4.6.1	Proposed CFCCIF-ESA Feature Set	101
4.6.2	Analysis of Phase-Related Artifacts in SS and VC SpooF . . .	107
4.6.3	Experimental Setup	109
4.6.4	Experimental Results on ASVSpooF 2015 Dataset	111
4.6.4.1	Comparison with Other Feature Sets on Eval Set .	111
4.6.4.2	Detailed Analysis	115
4.6.4.3	Assessment of the Proposed CFCCIF-ESA Feature Set using Various Performance Metrics	118
4.6.4.4	Performance Analysis on S10 Spoofing Attack . . .	120
4.7	Chapter Summary	122
5	Spectral-Based Features for Anti-spoofing	125
5.1	Introduction	125
5.2	CQT for Voice Liveness Detection (VLD)	126
5.2.1	Literature in Brief	126
5.2.2	VLD-ASV System and Baseline	129
5.2.2.1	Pop Noise and VLD System	130

5.2.2.2	STFT-Based Baseline Algorithm	131
5.2.3	Proposed CQT-Based Algorithm	134
5.2.3.1	Development of CQT	134
5.2.3.2	Spectrographic Analysis	137
5.2.3.3	CQT <i>vs.</i> STFT for Pop Noise Detection	138
5.2.4	Experimental Setup	143
5.2.4.1	Dataset Used	143
5.2.4.2	Classifiers Used	147
5.2.5	Experimental Results	147
5.2.5.1	Effect of Variation in Frequency Range	147
5.2.5.2	Effect of Number of Frames	149
5.2.5.3	Effect of Various Analysis Window Functions in CQT	150
5.2.5.4	Comparison of Results for STFT <i>vs.</i> CQT using Various Classifiers	151
5.2.5.5	Inclusion of Replay Mechanism	155
5.2.5.6	Performance Evaluation using ASVSpooF Challenge Datasets	160
5.3	Spectral Root Homomorphic Filtering-Based Features for Replay SSD in ASV and VAs	162
5.3.1	Speech Signal Modeling	163
5.3.2	Cepstrum Analysis: Logarithmic <i>vs.</i> Spectral Root	163
5.3.3	Proposed SRCC Feature Set	165
5.3.4	Experimental Setup	167
5.3.5	Experimental Results on ReMASC Dataset	167
5.3.6	Experimental Results on ASVSpooF 2017 Dataset	170
5.4	Chapter Summary	171
6	Other Applications	175
6.1	Introduction	175
6.2	Significance of CMVN for Replay SpooF Detection	176
6.2.1	Motivation	176
6.2.2	Cepstral Mean and Variance Normalization (CMVN)	177
6.2.3	Replay Speech Signal Modelling and CMVN	178
6.2.4	Experimental Setup	181
6.2.5	Experimental Results for ASVSpooF 2017 Dataset	182
6.2.6	Experimental Results for ASVSpooF 2019 dataset	185
6.3	DAS <i>vs.</i> MVDR Beamformer: Analysis for Replay SSD Task	189
6.3.1	Motivation	189

6.3.2	Signal Modeling for Microphone Array Signal	190
6.3.3	Delay and Sum (DAS) Beamformer	191
6.3.4	Minimum Variance Distortionless Response (MVDR)	193
6.3.5	Reverberation Analysis Using TEO	194
6.3.6	Experimental Setup	195
6.3.7	Experimental Results	195
6.4	Severity-Level Classification of Dysarthric Speech	198
6.4.1	Motivation	198
6.4.2	Problem Formulation	198
6.4.3	Characterizing Dysarthria in Speech Signals	200
6.4.4	Proposed Approach	202
6.4.4.1	Onset-Offset Detection	202
6.4.4.2	Spectrogram: T-F Representation	203
6.4.4.3	Mapping Technique: CNN <i>vs.</i> ResNet	203
6.4.5	Experimental Setup and Results	204
6.4.5.1	Dataset	205
6.4.5.2	Comparison Methods	205
6.4.5.3	Performance Evaluation	206
6.4.5.4	Analysis of Results	208
6.4.5.5	Complementary Comments	211
6.5	Classification of Normal <i>vs.</i> Pathological Infant Cry	211
6.5.1	Form-Invariance Property of CQT	212
6.5.2	Subband Teager Energy Representations	213
6.5.3	Experimental Setup	214
6.5.4	Experimental Results using CQCC	215
6.5.4.1	Spectrographic Analysis	215
6.5.4.2	Results using Evaluation Metrics	215
6.5.5	Experimental Results using Subband-TE Features	218
6.5.5.1	Spectrographic Analysis	218
6.5.5.2	Results using Evaluation Metrics	220
6.6	Chapter Summary	221
7	Summary and Conclusions	225
7.1	Summary of the Thesis	225
7.2	Limitations and Future Research Directions	228
7.3	Open Research Problems	231
Appendix A Heisenberg's Uncertainty Principle in Signal Processing Frame-		
work		233

Appendix B IF Estimation using Hilbert Transform 235
Appendix C IF Estimation using ESA 237
Appendix D Cauchy-Schwarz Inequality for Multichannel Noise Power 239
Bibliography 243
List of Publications from the Thesis 270

Abstract

Amongst various biometrics, voice is the most natural and convenient way of the communication for human-machine interaction. To that effect, the use of Automatic Speaker Verification (ASV) for authentication is increasing in various sensitive applications, which create a chance for fraudulent attack as attackers can breach the authentication by using various spoofing attacks. To alleviate this issue, we can either develop an ASV system, which is inherently protected from the spoofing attacks or develop a separate countermeasure (CM) system that can assist the ASV system in *tandem* against the spoofing attacks. The earlier approaches have trade-off between performance of the ASV system and robustness against spoofing attacks. Hence, it would be advantageous to implement the separate Spoof Speech Detection (SSD) system, and hence majority research attempts are focusing upon the later approach. To that effect, various international challenge campaigns were organized during INTERSPEECH conferences, such as ASVSpooF 2015, ASVSpooF 2017, and ASVSpooF 2019, which provides standard datasets, protocol, and evaluation metrics. This thesis focuses on developing the handcrafted feature sets for CM systems against the spoofing attacks, namely, Speech Synthesis (SS), Voice Conversion (VC), and replay. These feature sets are either developed by applying the subband filtering on the speech signals or derived from the spectrogram representations.

In this thesis work, various subband filtering-based feature sets are developed, namely, Enhanced Teager Energy-Based Cepstral Coefficients (ETECC), Cross-Teager Energy Cepstral Coefficients (CTECC), and Energy Separation Algorithm-based Instantaneous Frequency estimation for Cochlear Cepstral Features (CFCCIF-ESA). These feature sets are either modification in Teager Energy Operator (TEO)-based representations or utilization of Energy Separation Algorithm (ESA) for Instantaneous Frequency (IF) estimation. The ETECC feature set is developed by accurately estimating the energies in high frequency regions using compensation of the *signal mass*. In Teager Energy-Based Cepstral Coefficients (TECC), TEO is utilized to estimate the energy, which considers the approximation $\sin(\omega) \approx \omega$, which is applicable for low frequencies. However, the discriminative information

for the replay detection is prominently present in the mid and high frequency regions. Hence, ETECC feature set is proposed to obtain the efficient representation for SSD task by accurately estimating the energies at high frequency regions. Furthermore, signal processing-based approach is presented for replay SSD in Voice Assistants (VAs). It utilizes the Cross-Teager Energy Operator (CTEO) for extracting the acoustic cues from replay speech. CTEO gives the interactions among the multi-channel signal by estimating the cross-Teager energies between signals. To that effect, it is necessary to efficiently represent the acoustic cues for replay spoofs and hence, maximum cross-Teager energies among the subband filtered multi-channel signal is utilized for feature representation. Thus, the rationale behind optimal channel selection is to find the most noisy (distorted) transmission channel. The cepstral features extracted using CTEO are referred as Cross-Teager Energy Cepstral Coefficients (CTECC_{max}). The experiments are performed using *Realistic Replay Attack Microphone Array Speech Corpus* (ReMASC), which is specially designed for the replay SSD in VAs. The proposed CTECC_{max} feature set performs better than other state-of-the-art feature sets. The proposed CFCCIF-ESA feature set combines the magnitude and phase (in the form of IFs) information to develop the efficient feature representation for SS, VC, and replay spoofing attacks. The proposed CFCCIF-ESA utilizes ESA to accurately estimate the modulation patterns due to their relatively low computational complexity, high time resolution, and instantaneously adapting nature. In previously proposed Cochlear Filter Cepstral Coefficient Instantaneous Frequency (CFCCIF) feature set, IFs were estimated using Hilbert transform-based approach, whose time resolution is relatively low (as it requires a segment of speech) as compared to the ESA-based approach.

Furthermore, Constant-Q Transform (CQT)-based feature representation and Spectral Root Cepstral Coefficients (SRCC) are developed using spectrogram representations and effectively utilized for anti-spoofing. According to Heisenberg's uncertainty principle in signal processing framework, the CQT has variable spectro-temporal resolution, in particular, better frequency resolution for low frequency region and better temporal resolution for high frequency region. This property of the CQT representation is effectively utilized to identify the low frequency characteristics of pop noise. Here, pop noise is attributed to the live speaker and hence, it is exploited for Voice Liveness Detection (VLD) task. SRCC feature set is derived from the theory of homomorphic filtering, which obeys the generalized superposition theory. In spectral root homomorphic deconvolution system, convolutionally combined vectors are mapped to another convolutionally combined vector

space, where signal components are more easily separable by liftering operation. Logarithm operation in Mel Frequency Cepstral Coefficients (MFCC) extraction is replaced by power-law nonlinearity (i.e., $(\cdot)^\gamma$) to derive SRCC feature set. The proper choice of the γ depends upon the pole-zero arrangements in the transfer function obtained from the speech signal and it helps to capture the system information of the speech signal, with a minimum number of cepstral coefficients. In this thesis, optimum γ -value is chosen by estimating the energy concentration in cepstral coefficients and by visualizing the spectrogram w.r.t. γ -value.

To validate performance of our proposed feature sets, the experiments are performed using various datasets, state-of-the-art feature sets, classifiers, and evaluation metrics. The development and performance analysis of each proposed feature set is provided in the corresponding chapters. Furthermore, other contributions in the thesis, namely, feature normalization for anti-spoofing, analysis on Delay and Sum (DAS) *vs.* Minimum Variance Distortionless Response (MVDR) beamforming techniques for anti-spoofing in VAs, severity-level classification of dysarthric speech, and classification for normal *vs.* pathological cries, are also discussed. Thesis concludes with potential future research directions and open research problems.

List of Acronyms

ADAM	Adaptive Moment
ADC	Analog-to-Digital Converter
AEER	Average Equal Error Rate
AFCC	Adaptive Frequency Cepstral Coefficients
AIF	Average Instantaneous Frequency
AM	Amplitude Modulation
APGDF	All-pole Group Delay Function
ARP	Adaptive Relative Phase
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
AT	Auditory Transform
AUC	Area Under the Curve
BIBO	Bounded Input Bounded Output
Bi-LSTM	Bidirectional Long Short-Term Memory
BM	Basilar Membrane
BU	Butterfly Unit
CFCC	Cochlear Filter Cepstral Coefficient
CFCCIF	Cochlear Filter Cepstral Coefficient Instantaneous Frequency
CFCCIF-ESA	Energy Separation Algorithm-based Instantaneous Frequency estimation for Cochlear Cepstral Features
CGCNN	Context Gate CNN
CMC	Constant-Q Multi-level Coefficients
CMN	Cepstral Mean Normalization
CMVN	Cepstral Mean and Variance Normalization
CMS	Cepstral Mean Subtraction
CNN	Convolutional Neural Networks

CNPCC	Cosine Normalized Phase-based Cepstral Coefficients
CNPF	Cosine Normalized Phase Feature
ConvRBM	Convolutional Restricted Boltzmann Machine
CQCC	constant-Q Cepstral Coefficients
CQT	Constant-Q Transform
CQHC	constant-Q Harmonic Coefficients
CQSPIC	Constant-Q Statistics-plus-Principal Information Coefficients
CQ-EST	Constant-Q Equal Subband Transform
CQ-OST	Constant-Q Octave Sub-band Transform
CTE	Cross-Teager Energy
CTECC	Cross-Teager Energy Cepstral Coefficients
CTFT	Continuous-Time Fourier Transform
CTEO	Cross-Teager Energy Operator
DAS	Delay and Sum
DCF	Detection Cost Function
DCT	Discrete Cosine Transform
DET	Detection Error Trade-off
DESA	Discrete-time Energy Separation Algorithm
Dev	Development
DF-MST	Discrete Fourier Mel Subband Transform
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DSR	Distant Speech Recognition
DTFT	Discrete-Time Fourier Transform
DTFE	Direct Time Fundamental Frequency Estimation
EER	Equal Error Rate
EM	Expectation Maximization
ERB	Equivalent Rectangular Bandwidth
ESA	Energy Separation Algorithm
ESD	Energy Spectral Density
ETECC	Enhanced Teager Energy-Based Cepstral Coefficients
ETEO	Enhanced Teager Energy Operator
Eval	Evaluation
FC	Fully Connected

FFT	Fast Fourier Transform
FT	Fourier Transform
FFV	Fundamental Frequency Variation
FM	Frequency Modulation
FPI	Fullband Principal Information
GAT	Graph Attention Network
GD	Group Delay
GFCC	Gammatone Frequency Cepstral Coefficients
GLDS-SVM	Generalized Linear Discriminant Kernel
GMM	Gaussian Mixture Model
HOSP	Higher-Order Statistics Pooling
HT	Hilbert Transform
IBM	Ideal Binary Mask
ICASSP	International Conference on Acoustics, Speech, and Signal Processing
IF	Instantaneous Frequency
IFCC	Instantaneous Frequency Cepstral Coefficients
IMFCC	Inverse-MFCC
IMOBT	Inverted Mel Warping Overlapped Block Transformation
ISFCC	Inverted Signal-based Frequency Cepstral Coefficients
LA	Logical Access
LC-GRNN	Light Convolutional Gated Recurrent Neural Network
LCNN	Light-CNN
LFS	Linear Filterbank Slope
LFCC	Linear Frequency Cepstral Coefficients
LHDS	Logarithmic Homomorphic Deconvolution System
LinFB	Linear Filterbank Coefficients
LLR	Log-Likelihood Ratio
LMS	Log Mel Spectrogram
LPCC	Linear Prediction Cepstral Coefficients
LP	Linear Prediction
LTI	Linear Time-Invariant
M2VoC	Multi-speaker Multi-style Voice Cloning Challenge
MES	Maximum Energy Signal
MelFB	Mel Filterbank coefficients

MFCC	Mel Frequency Cepstral Coefficients
MFM	Mel Frequency Magnitude, Max-Feature-Map
MFP	Mel Frequency Phase
MFS	Mel Filterbank Slope
MGDF	Modified Group Delay Functions
MGDCC	Modified Group Delay Cepstral Coefficients
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MLT	Multilevel Transform
MLSA	Mel Log Spectrum Approximation
MOBT	Mel Warping Overlapped Block Transformation
MRCG	Multi-Resolution Cochleogram
MRP	Modified Relative Phase
MSRCC	Magnitude-SRCC
MTL	Multitask Learning
MVDR	Minimum Variance Distortionless Response
NN	Neural Network
NSD	Nerve Spike Density
NULBP	Normalized Unique Local Binary Patterns
OPI	Octave-band Principal Information
PA	Physical Access
<i>pdf</i>	Probability Density Function
PFE	Paraconsistent Feature Engineering
PLP	Perceptual Linear Prediction
PNCC	Power Normalized Cepstral Coefficients
POCO	POp noise COrpus
PSTN	Public Switched Telephone Network
PSRCC	Phase-SRCC
RASTA-PLP	Relative Spectral-PLP
ReMASC	Realistic Replay Attack Microphone Array Speech Corpus
ResNet	Residual Networks
RFCC	Rectangular Filter Cepstral Coefficients
RMS	Root Mean Square
RNN	Recurrent Neural Networks

RPS	Relative Phase Shift
RT	Reverberation Time
SCC	Scattering Cepstral Coefficients
SCFC	Subband Centroid Frequency Coefficients
SCMC	Spectral Centroid Magnitude Coefficients
SE	Squeeze-and-Excitation
SECC	Squared Energy Cepstral Coefficients
SFCC	Single Frequency Cepstral Coefficients
SFCC	Signal-based Frequency Cepstral Coefficients
SGD	Stochastic Gradient Descent
SID	Speaker Identification
SIDS	Sudden Infant Death Syndrome
SNR	Signal-to-Noise Ratio
SOBT	Signal-based Overlapped Block Transformation
SoE	Strength of Excitation
SRCC	Spectral Root Cepstral Coefficients
SRHDS	Spectral Root Homomorphic Deconvolution System
SSFC	Subband Spectral Flux Coefficients
SSD	Spoof Speech Detection
SS	Speech Synthesis
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation weiGHted specTrum
STFT	Short-Time Fourier Transform
STSSI	Short-Term Spectral Statistics Information
subband-TE	subband-Teager Energy representations
SVM	Support Vector Machine
TBP	Time-Bandwidth Product
TDNN	Time Delay Neural Networks
t-DCF	tandem - Detection Cost Function
TECC	Teager Energy Cepstral Coefficients
TTS	Text-to-Speech
UA	Universal Access
USS	Unit Selection Synthesis
VAs	Voice Assistants

VAE	Variational Autoencoder
VC	Voice Conversion
VCTK	Voice Cloning Toolkit
VLD	Voice Liveness Detection
VoIP	Voice-over-IP
VSDC	Voice Spoofing Detection Corpus
ZTWCC	Zero-Time Windowing Cepstral Coefficients

List of Symbols

$p(\cdot)$	Probability density function
$P(\cdot)$	Probability of an event
t	Time
ω	Frequency, wight vector
$x(t)$	Speech Signal
$x(n)$	Discrete-time Speech Signal
$X(z)$	System Function in Z-domain
$X(e^{j\omega})$	Frequency Response of System
$\psi\{\cdot\}$	Teager Energy Operator
$a_i(n)$	Instantaneous Amplitude
ω_i	Instantaneous Frequency
Σ	Notation for Summation
\int	Notation for integration
β	Fusion Parameter
$\phi(t)$	Instantaneous Phase
$*$	Convolution Operation
\approx	Approximately
Δ	Delta or Dynamic Features
$\Delta\Delta$	Delta-Delta or Double Delta or Acceleration Features
$E[\cdot]$	Expectation Operator
C^∞	Space of Infinitely Differentiable Functions
$L^2(\mathcal{R})$	Hilbert Space of Square Integrable Functions
\arcsin	Inverse Sine Function
\arctan	Inverse Tangent Function
γ	SRCC feature parameter,
F_0	fundamental frequency

μ	mean
σ	variance
λ_n	GMM model for natural speech
$\psi(\cdot)$	Feature space
ρ	Signal mass
$\rho(\cdot)$	Softmax function
Γ	Noise power

List of Tables

2.1	Results (in % EER and % AEER) from the Literature for Various SSD Systems on ASVSpooof 2015 challenge Dataset. After [1].	16
2.2	Results Obtained on ASVSpooof 2017 Version 2.0 Dataset for Various Systems Reported in the Literature.	19
2.3	Results Obtained on ASVSpooof 2019 Dataset (LA and PA Scenario) for the Various Architectures in the Literature.	20
3.1	Statistics of the ASVSpooof 2015 Challenge Dataset Partition. After [2].	27
3.2	Spoofing Algorithms Implemented in the ASVSpooof 2015 Challenge Dataset. After [2].	27
3.3	Statistics of the ASVSpooof 2017 Dataset for the Environment-Independent Case. After [3].	28
3.4	Distribution of Spoof Speech Utterances Among the Environments in ASVSpooof 2017 Dataset	28
3.5	Statistics of the ASVSpooof 2019 Dataset. After [4].	30
3.6	Algorithms for LA Spoofing Systems. Here, * Indicates Neural Network-Based Algorithm. After [4].	31
3.7	Parameter Settings for Acoustic Configurations to Generate Simulated Replay Spoofs ASVSpooof 2019 Challenge Dataset. After [5]. .	32
3.8	Parameter Settings for Replay Configurations in ASVSpooof 2019 Challenge Dataset. After [5].	32
3.9	Statistics of the BTAS 2016 Dataset <i>w.r.t.</i> the Session and Recording Type. After [6].	34

3.10	Number of Utterances in BTAS 2016 Dataset. Acronyms in this Table Stands for the Following Terms: SS- Speech Synthesis, VC- Voice Conversion, RE- Replay, LP- Laptop, PH1- Samsung Galaxy S4 Phone, PH2- iPhone 3GS, PH3- iPhone 6S, HQ- High Quality Speakers. After [7].	34
3.11	The Three Subsets of POCO Dataset. After [8].	35
3.12	The Set of Words Utilized in POCO Dataset. After [8].	36
3.13	Microphone Array Settings for ReMASC Dataset. After [9].	36
3.14	Statistics of the ReMASC Dataset <i>w.r.t.</i> Various Acoustic Environments. After [9].	36
3.15	Statistics of the Subset of the ReMASC Dataset Partitioned into Three Subsets. After [9].	37
3.16	Summary of Various Datasets Utilized in this Thesis.	38
4.1	Details of the Proposed CNN Architecture for SSD System. After [10].	67
4.2	Details of the Proposed LCNN Architecture for SSD System. After [10].	68
4.3	Evaluation of Various Feature Sets with Paraconsistent Framework on ASVSpooof 2017 Version-2 Dataset. After [10].	71
4.4	Results (in % EER) for ETECC Feature Set <i>w.r.t.</i> Type of the Filterbank. After [10].	73
4.5	Results (in % EER) for Varying Number of Mixtures in GMM for the Features Extracted using 40 Subband Filters in Gabor Filterbank. After [10].	76
4.6	Results (in % EER) for Subband Filters in Low, Mid, and High Frequency Regions. After [10].	77
4.7	Results (in % EER) for Static, Δ , and $\Delta\Delta$ Features for ETECC Feature Set. After [10].	78
4.8	Results (in % EER) on Eval and Dev Datasets for Individual SSD Systems. After [10].	79
4.9	Results (in % EER) on Score-Level Fusion on Eval and Dev Datasets using Linear and Logistic Regression. After [10].	79
4.10	Results (in % EER) of Environment-Dependent Case on ASVSpooof 2017 Dataset. After [10].	81
4.11	Results (in % EER) on ASVSpooof 2019 Challenge Dataset for PA Scenario. After [10].	83
4.12	Results (in % EER) on ASVSpooof 2019 Challenge Dataset for LA Scenario. After [10].	83

4.13	Results (in % EER) on BTAS Dataset. After [10].	83
4.14	Results (in % EER) on ASVSpooF 2015 Challenge Dataset. After [10].	83
4.15	Results (in % EER) on ReMASC Dataset using GMM Classifier. After [10].	83
4.16	Results (in % EER) for Environment-Dependent <i>vs.</i> -Independent Case on ReMASC Dataset on GMM Classifier. After [10].	84
4.17	Results (in % EER) on ReMASC Dataset. After [11].	95
4.18	Results in % EER for Environment-Dependent <i>vs.</i> Environment-Independent Case on ReMASC Dataset. After [11].	96
4.19	Results (in % EER) for Comparison of CTECC _{max} with Existing Architecture in [12] on Eval set for Various Devices. After [13].	99
4.20	Results (in % EER) on Dev and Eval Set <i>w.r.t.</i> Various Feature Sets and Devices using GMM Classifier. After [13].	99
4.21	Results (in % EER) on Dev and Eval Set <i>w.r.t.</i> Various Feature Sets and Devices using CNN Classifier. After [13].	100
4.22	Results (in % EER) on Dev and Eval Set <i>w.r.t.</i> Various Feature Sets and Devices using LCNN Classifier. After [13].	100
4.23	Details of the Proposed CNN Architecture for SSD System. After [1].	111
4.24	Results (in % EER and % AEER) from ASVSpooF Literature for Various SSD Systems Trained using GMM/SVM. The Performance of the Proposed Feature Set is Compared Against the Other Feature Sets in the Literature. After [1].	113
4.25	Results (in % EER and % AEER) from ASVSpooF Literature for Various SSD Systems Trained using DNN architectures. The Performance of the Proposed Feature Set is Compared Against the Other Feature Sets in the Literature. After [1].	114
4.26	Results (in % EER) on Proposed CFCCIF-ESA Feature Set Framework with Various Filterbank Structures. After [1].	117
4.27	Results in % EER, % Classification Accuracy, and AUC of Intersection of the Probability Density Functions (<i>pdfs</i>) Obtained from the LLR Scores for Dev and Eval Set of ASVSpooF 2015 Dataset. After [1].	118
4.28	Results (in % EER) for the Various Feature Sets for S10 Spoofing Attack Detection. After [1].	120
5.1	The Comparison of the CQT, <i>Resampled</i> CQT, and STFT <i>w.r.t.</i> the Various Spectrographic Parameters for Pop Noise Detection with $F_s = 22050$ Hz. After [14].	141

5.2	Window Length in Samples as a Function of Analysis Frequency (f_k). After [14].	143
5.3	Statistics of the POCO Dataset used for Experiments in this Thesis. After [8,15].	145
5.4	Parameter and Corresponding Configuration for Replay Mechanism. After [16].	146
5.5	Details of the Proposed LCNN Architecture for VLD. After [17]. . .	148
5.6	Details of the Proposed ResNet Architecture for VLD. After [18]. . .	149
5.7	Results (in % Classification Accuracy) for CQT-SVM-based Pop Noise Detection using RC-A (genuine) <i>vs.</i> RP-A (spoof) Dataset with Variation in Frequency Range. After [15].	149
5.8	Results in (% Classification Accuracy) for the Original CQT-based Algorithm <i>vs.</i> Resampled CQT-based Algorithm using SVM Classifier on POCO Dataset. After [15].	149
5.9	Results (in % Classification Accuracy) for Varying the Number of Frames in Proposed CQT-based Algorithm with SVM Classifier on POCO Dataset.	150
5.10	Results (in % Classification Accuracy) of Proposed CQT-based Approach with Different Window Functions using Various Classifiers.	151
5.11	Comparison of Proposed CQT-based Approach with the STFT-based Baseline Approach using Various Classifiers. After [15].	152
5.12	Comparison of CQCC and LFCC Feature Sets using Various Classifiers on POCO Dataset. After [15].	153
5.13	Comparison of Proposed Approach <i>vs.</i> the Baseline Approach with Larger Training Data (80 % Training, 20 % Testing) for Various Classifiers on POCO Dataset. After [15].	155
5.14	Comparison of Baseline <i>vs.</i> Proposed Approach for Different Types of Phonemes using Various Classifiers. After [15].	158
5.15	Effect of Varying Distance between Subject Speaker and Microphone on Performance (in % Classification Accuracy) of RC-A <i>vs.</i> REP-A Subset with SVM as a Classifier. After [15].	159
5.16	Effect of Varying Distance between Subject Speaker and Microphone on Performance (in % Classification Accuracy) of RP-A <i>vs.</i> REP-A Subset with SVM as a Classifier. After [15].	160
5.17	Results (in % Classification Accuracy and % EER) for RC-A <i>vs.</i> REP-A with Various Frequency Ranges. After [15].	161

5.18	Results (in % EER) on ASVSpooF 2019 PA and ASVSpooF 2017 Version-2.0 Dataset using Proposed CQT-based Feature Set <i>vs.</i> CQCC (Challenge Baseline).	161
5.19	Variation in % EER <i>w.r.t.</i> γ Value for MSRCC Feature Set. After [19].	169
5.20	Results (in % EER) on ReMASC Dataset using Various Feature Sets. After [19].	170
5.21	Results (in % EER) on ASVSpooF 2017 Dataset. After [20].	171
6.1	Results of CQCC-GMM and LFCC-GMM Systems in % EER for Environment-Independent Case on ASVSpooF 2017 Dataset. After [21].	184
6.2	Results of CQCC-GMM System in % EER for Environment-Dependent Case on ASVSpooF 2017 Dataset. After [21].	185
6.3	Results of CQCC-GMM Systems ASVSpooF 2019 Dataset using Standard Protocols.	186
6.4	Results (in % EER) for Environment-Dependent Case for Various Replay Configurations on ASVSpooF 2019 Challenge Dataset. . . .	188
6.5	Results in (% EER) for Environment-Dependent Case for Various Acoustic Configurations on ASVSpooF 2019 Challenge Dataset. . .	189
6.6	Results (in % EER) on ReMASC and its DAS <i>vs.</i> MVDR Beamformed Versions using Various Feature Sets and Classifiers. After [22].	197
6.7	Severity-Level Classification Based on Intelligibility. Adapted from [23].	199
6.8	Proposed Architectural Details of ResNet. Here, Conv1 and Conv2 show Continuous Layers of Residual Block, and Conv3 shows Parallel Downsampling Layer in Residual Block. After [24].	205
6.9	Details of the Proposed LCNN Architecture for Dysarthria Severity Classes. After [24].	207
6.10	Evaluation of Baseline CNN <i>vs.</i> ResNet When Entire Speech Utterance is Available for Training. After [24].	210
6.11	Statistics of the Baby Chillanto dataset. After [25].	214
6.12	Results in % Classification Accuracy (Acc) for Various f_{min} (Hz) of using GMM. After [26].	215
6.13	Results (in % Classification Accuracy) for Various Window Functions using GMM. After [26].	217
6.14	Results (in % Classification Accuracy) <i>w.r.t.</i> Number of Mixtures. After [26].	217

6.15	Results in (% Classification Accuracy and % EER) for Various Feature Sets using GMM as a Classifier. After [26].	218
6.16	Results in (% Classification Accuracy and % EER) using Various Cepstral Feature Sets using GMM and SVM as Classifiers. After [27].	219
6.17	Results in (% Classification Accuracy and % EER) for Various Spectral Feature Sets using GMM and SVM as Classifiers. After [27]. . .	219
6.18	Results (in % Classification Accuracy) <i>w.r.t.</i> Number of Mixtures. After [27].	221
6.19	Results in % Classification Accuracy (Acc) for Various Number of Filters using GMM. After [27].	221

List of Figures

1.1	Illustration of the Typical ASV System with Possible Spoofing Attack Points. Point-1 and Point-2 Corresponds to Direct Attacks. Point-3 to Point-8 Corresponds to Indirect Attacks. Adapted from [28].	3
1.2	Functional Schematic of SSD System in <i>Tandem</i> with ASV System. After [29].	5
1.3	Flowchart Depicting Organization of this Thesis.	10
3.1	Residual Learning: A Building Block. After [18].	43
3.2	A Tandem System Consisting of ASV and SSD Modules is Evaluated using Three types of Trials: Targets, Nontargets, and Spoofing Attacks. Adapted from [30].	46
3.3	Demonstration of the AUC of the Overlapping Regions for the <i>pdfs</i> of the LLR Scores for the Two-class Classification Task. After [1]. . .	47
4.1	Brief Illustration of the Chronological Development of the Proposed TEO-based Features: ETECC, CTECC, and CFCCIF-ESA.	52
4.2	(a) A Synthetic AM-FM Signal Along With Superimposed Signal Mass, and (b) TEO, ETEO Profile Along With Signal Mass for the Signal Shown in Figure 4.2-(a). After [10,31].	58
4.3	Functional Block Diagram of the Proposed ETECC Feature Set. After [32].	59
4.4	(a) Impulse Response of 2^{nd} a Bandpass Filter in the Bank of 10 Subband Filters in a Gabor Filterbank with Linearly-Spaced Center Frequencies Between 0 Hz to 8 kHz, and (b) Frequency Response of a Gabor Filterbank.	62
4.5	Energy Estimated by TEO and ETEO for (a) 5^{th} and (c) 15^{th} Subband Filter Output. A Magnified View for the Corresponding Encircled Region is Shown in (b), and (d), Respectively.	64

4.6	Waterfall Plot of <i>Genuine</i> (Panel I) and <i>Spoof</i> (Panel II) Speech Utterances: (a), (c) Waterfall Plots Obtained using TECC Feature Set, and (b),(d) Waterfall Plots Obtained using ETECC Feature Set. After [32].	65
4.7	Frequency Response of (a) Gabor Filterbank, (b) Cochlear Filterbank, (c) Gammatone Filterbank, and (d) Mexican-hat Filterbank. .	72
4.8	ETEO-spectrogram Representation Obtained from (a) Gabor, (b) Cochlear, (c) Gammatone, and (d) Mexican-hat Filterbank with 40 Subband Filters in the Filterbank for Genuine (Panel-1), and it's Corresponding Spoof (Panel-2) Speech Utterance.	72
4.9	Variation of % EER of Dev set with the (a) Bandwidth of Subband Filters in the Gabor Filterbank, and (b) Speech Frame Length of Analysis Window during ETECC Feature Extraction. After [10]. . .	74
4.10	Variation of % EER of Dev Set with the (a) Number of Subband Filters, and (b) Dimensions of ETECC Feature Vector (Including Static, Δ , and $\Delta\Delta$ Coefficients). After [10].	75
4.11	Individual DET Curves on (a) Dev, and (b) Eval Sets. After [10]. . .	80
4.12	LLR Scores Distribution on Dev Set, (Panel I) and Eval Set (Panel II) using (a),(b) CQCC, (c),(d) TECC, and (e),(f) ETECC Feature Sets. After [10].	80
4.13	Selected Chronological Progress to Develop CTECC _{max} Feature Set for Anti-Spoofing.	86
4.14	Functional Block Diagram of Proposed CTECC _{max} Feature Extraction. After [11,13].	91
4.15	Spectrogram Plot of CTECC _{max} (Panel I) <i>vs.</i> CQCC (Panel II) Feature Sets : (a), (b) for Genuine Speech Signal, and (c), (d) for Spoofed Speech Signal. After [11].	94
4.16	Results using CTECC _{max} <i>w.r.t</i> Number of Subband Filters used in Gabor Filterbank: (a) Dev Set, and (b) the Eval Set.	97
4.17	Results <i>w.r.t</i> Dimension of CTECC _{max} Feature Vector: (a) Dev Set, and (b) Eval Set.	97
4.18	Results <i>w.r.t</i> Number of Mixtures in GMM using CTECC _{max} : (a) Dev Set, and (b) Eval Set.	98
4.19	LLR Scores Distribution on Eval Set of D4: (a) MFCC, (b) CQCC, (c) LFCC, (d) TECC, (e) CTECC _{min} , and (f) CTECC _{max} . After [13]. . . .	100

4.20	DET Curves Obtained for Various Feature Sets as Shown in Legends. Figure 4.20(a) and Figure 4.20(e) shows the DET Curves for Device D1 on Dev and Eval Set, Respectively. Similarly, (Figure 4.20(b), Figure 4.20(f)), (Figure 4.20(c), Figure 4.20(g)), and (Figure 4.20(d), Figure 4.20(h)) shows the DET Plots for Device D2, D3, and D4 on (Dev, Eval) Set, Respectively. The Legend shown in Figure 4.20(a) is Similar for Remaining DET Plots.	101
4.21	Functional Block Diagram of the CFCC, CFCCIF, and Proposed CFCCIF-ESA Feature Set. After [1,33].	102
4.22	Selected Chronological Progress of the Proposed CFCCIF-ESA Feature Set. After [1].	102
4.23	(a) Impulse Response of 2 nd Subband (Cochlear) Filter, and (b) Corresponding Frequency Response of Cochlear Filterbank Consisting of Ten Subband Filters. After [1].	103
4.24	Ten IF Contours of the Subband Filtered Genuine Speech Signal. Subband Filtering is Performed by the Cochlear Filterbank with Ten Subband Filters (and hence, Ten IF Contours) Covering the Nyquist Sampling Frequency Range. After [1].	110
4.25	Ten IF Contours of the Subband Filtered SS- and VC-based Spoof Speech Signal. Subband Filtering is Performed by the Cochlear Filterbank with Ten Subband Filters (and hence, Ten IF Contours) Covering the Nyquist Sampling Frequency Range. After [1].	110
4.26	Results (in % EER) <i>w.r.t.</i> Number of Subband Filters on the Dev Set. After [1].	116
4.27	Frequency Response of Filterbanks with 10 Subband Filters: (a) Gabor, (b) Cochlear ($\alpha = 3$ and $\beta = 0.005$), and (c) Gammatone. After [1].	117
4.28	LLR Score Distribution of Genuine <i>vs.</i> Spoof Speech Distribution for the SSD Systems Developed using Various Feature Sets and Classifiers. (Figure 4.28 (a) and (f)), (Figure 4.28 (b) and (g)), (Figure 4.28 (c) and (h)), (Figure 4.28 (d) and (i)), and (Figure 4.28 (e) and (j)) Shows the LLR Score Distribution for the SSD Systems CFCC-GMM, CFCCIF-GMM, CFCCIF-ESA-GMM (A), CFCCIF-ESA-CNN (B), Score-Level Fusion of (A) and (B) on Dev and Eval Set, Respectively. After [1].	119

4.29	DET Curves Obtained from the SSD Systems Implemented using CFCC, CFCCIF, and Proposed CFCCIF-ESA Feature Sets on ASVSpooof 2015 dataset. Figure 4.29(a) and Figure 4.29(b) Shows the DET Plots for Dev and Eval Set, Respectively. DET Curves for CFCCIF-ESA-CNN and CFCCIF-ESA-Fusion are not Visible in Figure 4.29(a) Due to % EER is Approaching to Zero. After [1].	119
4.30	The Speech Signal and Spectrographic Representation of the Genuine <i>vs.</i> S10-attack. Panel-I and Panel-II Shows the Speech Signal with its Spectrographic Representation of the Genuine and S10-attack, Respectively. Figure 4.30(a) Represents the Speech Signal. Whereas, Figure 4.30(b), Figure 4.30(c), and Figure 4.30(d) Represents the Corresponding Spectrogram Obtained from STFT, CQT, and CFCCIF-ESA Feature Set, Respectively. After [1].	121
5.1	Selected Chronological Progress for CQT and its Derived Feature Sets for Speech Analysis and Anti-spoofing Tasks. After [15].	128
5.2	A Schematic of VLD System in Tandem with ASV System. After [34].	130
5.3	Functional Block Diagram of Baseline and Proposed Algorithm. After [35].	133
5.4	Panel-I and Panel-II Depicts the Spectrographic Analysis for Genuine <i>vs.</i> Spoof Speech Signal for Word "Laugh", Respectively. (a) the Waterfall Plot for STFT, (b) the Top-view of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the Top-view of the CQT Waterfall Plot. The Rectangular Box Represents the Intended Location of the Pop Noise, whereas the Encircled Region Represents the Presence of F_0 and Its Harmonics. After [15].	139
5.5	Panel-I and Panel-II Depicts the Spectrographic Analysis for Genuine <i>vs.</i> Spoof Speech Signal for Word "Chip", Respectively. (a) the Waterfall Plot for STFT, (b) the Top-view of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the Top-view of the CQT Waterfall Plot. The Rectangular Box Represents the Intended Location of the Pop Noise, whereas the Encircled Region Represents the Presence of F_0 and Its Harmonics. After [15].	140
5.6	(a) Temporal Variance (σ_t^2), (b) Frequency Variance (σ_ω^2), and (c) TBP (i.e., $\sigma_t^2 \cdot \sigma_\omega^2$) for Hamming, Gaussian, and hann Windows. After [15].	144
5.7	Schematic Diagram for Generation of Synthetic Replay using Image Source Model. After [36].	146

5.8	DET Curves for the Proposed CQT-based Algorithm <i>vs.</i> STFT-based Baseline Algorithm for Various Classifiers on (a) Dev, and (b) Eval Set. Legends in Figure 5.8(b) are the Same as that of Figure 5.8(a). After [15].	154
5.9	Comparison of Wordwise % Classification Accuracy on Dev set with (a) SVM, (b) GMM, (c) CNN, (d) LCNN, and (e) ResNet as Classifier for STFT (Baseline) and CQT (Proposed) Feature Set. After [15].	156
5.10	Comparison of Wordwise % Classification Accuracy on Eval set with (a) SVM, (b) GMM, (c) CNN (d) LCNN, and (e) ResNet as Classifier for STFT (Baseline) and CQT (Proposed) Feature Set. After [15].	157
5.11	Functional Block Diagram of SRCC (MSRCC and PSRCC) Feature Set Extraction. After [20].	166
5.12	Panel-I and Panel-II Consists of Spectrogram of Genuine <i>vs.</i> Spoof Speech Signal, Respectively. Figure 5.12(a) Shows Spectrogram of the Speech Signal as Given in eq. (5.37) for $\gamma = 0.9$. Whereas, Figure 5.12(b), Figure 5.12(c), and Figure 5.12(d) Shows the Spectrogram for $\gamma = -0.9, 0.1$, and -0.1 , Respectively. After [19].	168
5.13	Plot of RSMS (Panel I) <i>vs.</i> CQT-gram (Panel II) Feature Sets : (a), (b) for Genuine Speech Signal, and (c), (d) for Spoofed Speech Signal. After [19].	169
5.14	DET Curves for (a) Dev Set and (b) Eval Set of ASVSpooof 2017 Challenge Dataset. After [20].	172
6.1	Scatter Plot for (a) the Unnormalized Data, (b) with CMN, and (c) CMVN. $X = [x_1 \ x_2]$ Denotes the Samples Drawn from the Bivariate Gaussian Distribution. Ytick Values of Figure 6.1(b) and Figure 6.1(c) are the Same as that of Figure 6.1(a). After [21].	178
6.2	Scatter Plot for (a) the Unnormalized Data, (b) with CMN, and (c) CMVN. $X = [x_1 \ x_2]$ Denotes the First and Second Dimension of CQCC Feature Vector. Legends of Figure 6.2(b) and Figure 6.2(c) are the Same as That of Figure 6.2(a). After [21].	180

6.3	Estimated <i>pdf</i> of Genuine and Environmentwise Spoof Speech Samples over the (a) 1 st , (b) 3 rd , (c) 5 th , (g) 10 th , (h) 12 th , (i) 15 th , (m) 20 th , (n) 25 th , and (o) 30 th Feature Dimensions with Application of CMVN, whereas Figure (d), (e), (f), (j), (k), (l), (p), (q), and (r) shows the Estimated <i>pdfs</i> for without CMVN Case with the Same Sequence of Dimensions as that of CMVN Case. Legends of all Figures Are Similar as Given in Figure 6.3(a). After [21].	182
6.4	Estimated <i>pdf</i> of Genuine and Spoof Speech Samples over the First Feature Dimension for (a) CMVN and (b) without CMVN. Legends of Figure 6.4(b) Are Similar to that of Figure 6.4(a). After [21]. . . .	183
6.5	DET Plots for Environment-Dependent Case using ASVspoof-2017 Dataset (a) with Application of the CMVN, and (b) without Application of the CMVN on Feature Set. Legends for Figure 6.5(a) and Figure 6.5(b) Are the Same. After [21].	186
6.6	All the <i>pdfs</i> Shown in Figure 6.6 Are Estimated from 1 st Cepstral Coefficient of the CQCC Feature Set. Figure 6.6(a) and Figure 6.6(c) Shows the Estimated <i>pdfs</i> for the Genuine <i>vs.</i> Spoof Speech Samples without Normalization, and CMVN, Respectively. Figure 6.6(b) Shows the Estimated <i>pdfs</i> for the Genuine <i>vs.</i> Individual Replay Configurations with CMVN Applied on CQCC Feature Set. Whereas, Figure 6.6(d), Figure 6.6(e), and Figure 6.6(f) Shows the Estimated <i>pdfs</i> for the Genuine <i>vs.</i> Individual Replay Configurations without Normalization Applied on CQCC Features. After [21].	187
6.7	Functional Block Diagram of DAS Beamformer having <i>N</i> Number of Microphones in an Array. After [37].	192
6.8	Time-domain Representation of (c) Genuine <i>vs.</i> (d) Replayed Speech Signal from ReMASC Dataset. Figure 6.8(a) and Figure 6.8(b) Represents the Zoomed Version of the Dotted Squared Region and Figure 6.8(e) and Figure 6.8(f) Corresponds to the Zoomed Version of the Solid Squared Region from Figure 6.8(c) and Figure 6.8(d), Respectively. After [22].	195
6.9	TEO Profile of Genuine (Panel I) and Replayed (Panel II) Speech Signals from (a) Original ReMASC and its (b) DAS <i>vs.</i> (c) MVDR Beamformed Versions. After [22].	196
6.10	DET Curves for ReMASC and its Beamformed Versions using TECC with GMM: (a) Dev, and (b) Eval set. After [22].	197

6.11	STFT Representation of: (a) Normal Speech, (b) Dysarthric Speech <i>vs.</i> LP Spectrum of (c) Normal Speech, and (d) Dysarthric Speech. After [24].	200
6.12	Waterfall Characteristics of: (a) Normal Speech, and (b) Dysarthric Speech. After [24].	201
6.13	Schematic Representation of F_0 Detection. Adapted from [38]. . . .	202
6.14	Schematic Representation of Proposed Methodology. After [39]. . .	202
6.15	STFT of the One Second of Speech Segment of Speakers with Differ- ent Severity-Levels, When They Pronounce the Word "Command": (a) Very Low, (b) Low, (c) Medium, and (d) High. After [24].	204
6.16	Baseline CNN <i>vs.</i> ResNet, for Different Speech Duration Based on (a) Classification Accuracy Score, and (b) F1-Score. Additionally, LNCC and GMM were also Considered for Comparisons, however, Since GMM Exhibit a Poor Accuracy, Its F1-Scores were Not Even Computed. After [24].	208
6.17	Learning of Proposed ResNet <i>vs.</i> Baseline CNN. For Both Panels, We Have: [First Column]: Input Spectrogram of Chunk (Horizontal- Axis: Time, Vertical-Axis: Frequency); [Second Column]: Visualiza- tion of Learning of ResNet; [Third Column]: Visualization of Learn- ing of CNN; [Forth Column]: Visualization of Learning of LCNN. Here, Visualization Images Are in the Form of Pixels. After [24]. . .	209
6.18	Evaluation of Baseline CNN <i>vs.</i> ResNet for Different Number of Chunks (i.e., Amount of Training Data) Based on (a) Classifica- tion Accuracy Score, and (b) F1-Score. As Previously Presented, GMM and LCNN were Considered for Comparisons, However, Since GMM Exhibit a Poor Accuracy, Its F1-scores Were Not Even Computed. After [24].	210
6.19	Functional Block Diagram of the Proposed Subband TEO Repre- sentation and TECC Feature Set. (SF: Subband Filtered Signal, TE: Teager Energies, AE: Averaged Energies over Frames). After [40,41].	214
6.20	Panel-I and Panel-II Depicts the Spectrographic Analysis for Healthy (Normal) and Pathology (Asphyxia) Infant Cry Signal: (a) the Wa- terfall Plot for STFT, (b) the Top View of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the Top View of the CQT Waterfall Plot. After [24].	216
6.21	DET Curves Obtained for Various Features using GMM and SVM Classifiers. After [26].	218

6.22	Panel-I and Panel-II Represents the Spectrographic Analysis for Normal <i>vs.</i> Asphyxia Cry Samples, Respectively. Figure 6.22(a), Figure 6.22(b), and Figure 6.22(c) Represents the STFT, MelFB, and Subband-TE Representations, Respectively. After [27].	219
6.23	DET Plots for Various Feature Sets using GMM and SVM as Classifiers. After [27].	220

CHAPTER 1

Introduction

1.1 Motivation

Speech is the primary and widely used mode of communication between humans. It is a form of quasi-periodic signal with a basic purpose of transmission (and reception) of information from one person to the other, or from a person to a machine or vice-versa [42]. Speech complexity can be appreciated by the fact that an utterance can convey completely different meanings simply by stressing a particular word or changing the position of pauses. Even though speech is a one-dimensional signal, it contains several levels of the information, such as linguistic message, speaker's identity, gender, health, emotion, attitude, acoustic environment, etc. [43]. This sophisticated information in the speech signal can be exploited for the advancement in technology to leverage the quality of human life. To that effect, several speech signal processing-based applications are developed, such as Automatic Speech Recognition (ASR), automatic speaker recognition, speech enhancement, speaker diarization, voice conversion, infant cry analysis, dysarthric speech classification and enhancement, whispered speech recognition, and many more. Among various speech technologies, an automatic speaker recognition system aims at identifying speaker-specific information, which can be used later for recognition purposes [44]. An automatic speaker recognition system can be divided into two categories, namely, Speaker IDentification (SID) and Automatic Speaker Verification (ASV). A SID system identifies an individual speaker from a pool of speakers, which may be an open or closed set of speakers [43,45]. An ASV system, on the other hand, verifies the claimed identity of the speaker [46].

Various biometric traits have been successfully deployed for different person verification systems, such as voice, signature, gait, face, iris, fingerprint, palmprint, palm/finger vein, etc. [47]. Naturally, humans have an ability to identify a person jointly using face and voice biometrics, i.e., through multimodal process-

ing. With the recent development in technology, ASV systems are effectively used as a voice biometric. The advantage of using speech as a biometric lies in its simplicity to gain a access remotely to recognition systems. Other biometrics, such as iris and thumb impression, require the person to be in physical contact with the system for the purpose of verification. However, voice biometric provides a touch-free verification environment. Recent developments in voice biometrics have got them wide acceptance because of their convenience, in particular, voice is natural to produce and non-invasive to collect via low cost microphones. Advancements in computational capabilities have led to better machine learning algorithms through which noise robust high performance ASV systems have been developed [45].

Figure 1.1 shows the illustration of a generic ASV system. A typical ASV system operates in two stages, namely, the enrollment stage and the recognition stage. In the enrollment stage, the speech samples are acquired from the speakers and a salient feature set (speaker-specific) extracted and stored in a database (often referred to as a template or speaker model), along with an identifier. During the verification stage, the system once again acquires the speech samples of a speaker, extracts a feature set from it, and compares this feature set against the templates in the database in order to verify the claimed identity. The comparison is performed by the *matcher* as shown in Figure 1.1. The matcher determines a match score, which represents the relative similarity of the sample to an already stored template. Finally, the decision module uses the relative score (usually, a log-likelihood ratio, i.e., LLR) to either accept or reject the claimed identity.

The desirable characteristics of the speaker-specific feature set for designing ASV system are [43,46]:

- Occurring naturally and frequently in speech.
- Easy to extract.
- Efficient in representing the speaker-dependent information and should not be affected by speaker's health and emotions.
- Stable over time.
- Not getting affected *w.r.t.* transmission channel characteristics or acoustic environments.
- Not susceptible to various spoofing attacks, such as identical twins, professional mimics, speech synthesis (SS), voice conversion (VC), and replay.

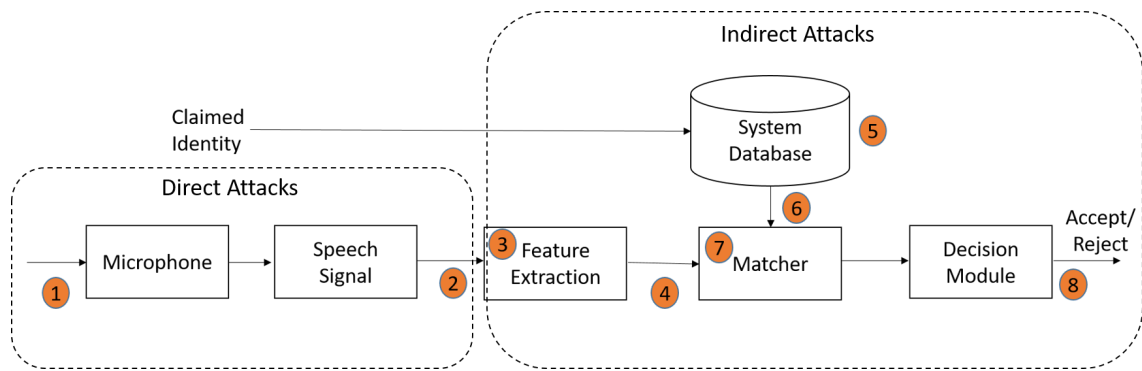


Figure 1.1: Illustration of the Typical ASV System with Possible Spoofing Attack Points. Point-1 and Point-2 Corresponds to Direct Attacks. Point-3 to Point-8 Corresponds to Indirect Attacks. Adapted from [28].

In general, the feature sets can be classified as handcrafted features *vs.* data-driven features. The extraction of crucial handcrafted features for speech applications requires knowledge of the speech production and perception mechanism. Considering this domain knowledge, mathematical models are defined based on the underlying physiological transformations in the speech production/perception mechanism. On the other hand, data-driven features extract the meaningful information mostly using unsupervised learning. It requires powerful deep learning algorithms, which should have the ability to learn features automatically. It also requires complex network architecture and a considerable amount of calculation time to attain better accuracy of classification. Furthermore, sufficient amount of statistically meaningful data is required for the training the deep learning algorithms, which can extract the meaningful features. However, handcrafted features can be employed for the intended task irrespective of the size of the dataset. This thesis is based on the development of the handcrafted features for the anti-spoofing task.

Advancements in the technology have led to increase in vulnerability to various spoofing attacks mentioned above [47]. Spoofing refers to an intentional circumvention, wherein an imposter tries to manipulate a biometric system simply by masquerading as another genuinely enrolled person [28,48]. The various components of ASV and links between them are vulnerable to possible attack points, as shown in Figure 1.1 [49]. These spoofing attacks can be categorized as *direct* attacks and *indirect* attacks. Direct attacks are applied at the microphone and transmission levels, which are labelled as point-1 and point-2, respectively, in Figure 1.1. These attacks include the impersonation (a.k.a. human mimicking), which can be performed by the identical twins and professional mimics by exploiting physiological characteristics and skillfulness, respectively. Furthermore, it also in-

cludes the SS, VC, and replay attacks, which can be presented at the microphone and transmission channel. Indirect attacks are performed within the ASV system itself, which are shown as attack points-3 to -8 in Figure 1.1. Indirect attacks generally require system-level access, for example attacks that interfere with feature extraction (point-3 and -4), models (point-5 and -6) or score and decision logic computation (point-7 and -8). This thesis mainly focuses on technology-based direct attacks, namely, SS, VC, and replay.

In the 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 toward the natural speech [50, 51]. Furthermore, there are advancement in SS- and VC-based technologies lead to vocoders, neural network-based generative architecture, such as WavNet, which can be used to embed the speaker-specific characteristics in the speech signals. In addition, replay signals are more threat to ASV systems as high quality recording devices are easily available in the market. Furthermore, Voice Assistants (VAs) are also emerging in recent days, which are utilized to control smart home appliances, activate home security systems, purchase items online, initiate phone calls, and complete many other tasks with ease. It is all the more hazardous if a fraudulent person could access the VAs using spoofing attacks. Hence, we would expect that spoofing attack should be alleviated for ASV and VAs. It can be accomplished by either developing the ASV or VA system, which has inherent capability to alleviate spoofing attacks [52] or implementing the separate countermeasure (CM) system against spoofing along with ASV system [28]. However, there is a trade-off between noise robust ASV system and its capability to resist the spoofing attacks. Hence, the latter approach of the developing separate CM system against spoofing attack is advantageous, and this thesis contributes to feature-based approach for developing CM systems.

The functional schematic of the Spoof Speech Detection (SSD) or CM system along with ASV system is shown in Figure 1.2. The ASV system verifies the claimed identity by using the speech signal presented to the ASV system. However, the presented signal can be a spoofed speech signal, and it can attack the ASV system if its characteristics are very similar to the stored templates (i.e., speaker models) in the ASV system. To alleviate this issue, the SSD system is employed as a CM system before ASV system as shown in Figure 1.2. The SSD system will identify whether the presented speech signal is coming from the natural source of the speech signal (i.e., natural speech production mechanism). If the presented

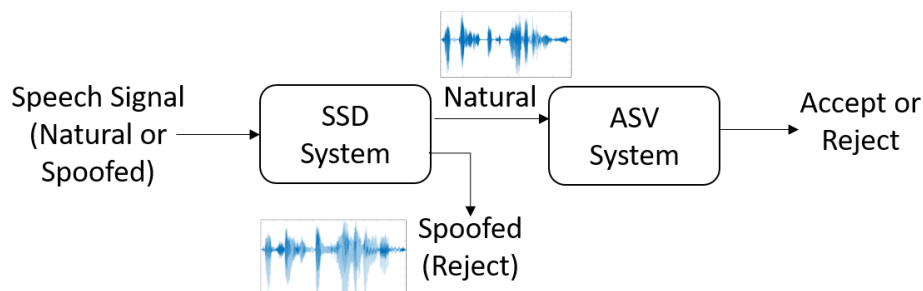


Figure 1.2: Functional Schematic of SSD System in *Tandem* with ASV System. After [29].

speech signal is identified with synthetic or replay speech characteristics, then it is considered as spoof speech signal, and it is restricted from the further processing.

1.2 Unnaturalness in the Spoof Speech Signals

Let us consider the natural speech production mechanism in brief, to understand the difference between naturally produced speech and spoof speech signals. Natural speech signal is produced using coordination among various speech organs, namely, lungs, larynx, and vocal tract system. Lungs act as a power supply and provides controlled airflow to the larynx stage. The airflow is modulated by the larynx and produces either periodic air puffs, noisy airflow or impulse-like excitation to the vocal tract system. These modulated signals are also considered as source signal in signal processing perspective. These source signals are provided to the vocal tract system, which mainly includes oral, nasal, and pharynx cavities. The vocal tract system is considered as system in order to model the speech production mechanism and hence, it supposed to perform the necessary *spectral coloring* of laryngeal excitation source signal. To implement the spoofing attack, the speech signal can be manipulated using SS, VC or replay mechanism. During this manipulation and implementation of the spoofing attacks, unnaturalness (or artifacts) is introduced in the spoof speech signal. The possible characteristics of this unnaturalness in the spoof signal are discussed next:

- Unnaturalness in SS and VC:
 - a. SS- and VC-based speech signals are conventionally generated using vocoders and unit selection synthesis (USS). Speech Transformation and Representation using Adaptive Interpolation weiGHted specTrum (STRAIGHT) and Mel log spectrum approximation (MLSA) vocoders

- are generally used to build SS- and VC-based spoofing attacks. Various factors may cause the quality degradation for the vocoder generated speech signals, such as modelling accuracy and quality of vocoders, oversmoothing of spectral parameter trajectories, overfitting, and time-independent mapping in VC techniques.
- b. Furthermore, vocoder-based signals are generated using magnitude information in the spectrum, ignoring the phase-based information. The artifacts generated due to these degradation factors can be utilized for the SSD task. In USS, which is utilized for speech synthesis, speech segments corresponding to speech sound units (such as, phonemes, syllables, etc.) are concatenated. The unnaturalness is introduced in the USS-based synthetic speech due to various factors, such as error in automated labelling of the speech sound units, the discontinuity at the joints, linear phase mismatches at the joints, lack of text-dependent prosody, etc.

However, LA scenario in ASVSpooF 2019 dataset consists of advanced architectures of Text-to-Speech (TTS) and VC, which produces the spoofed speech signals with high perceptual naturalness and speech quality. Some spoofed data is even challenging to detect for human beings. Therefore, the ASVspooF 2019 database is expected to be used to examine how CMs perform facing the advanced TTS and VC spoofing systems. This is also analyzed by visualizing features for spoofed *vs.* bonafide in a 2-D space followed by the clustering process to find out which attacks are naturally grouped together [4].

The speech signal which is secretly (and distantly) recorded from the genuine speaker and played back is known as replay speech signal. To mount the spoofing attack, the replay speech signal is presented to the ASV/VA system.

- Unnaturalness in Replay Speech Signal:
 - a. The additional processing for the generation of the replay speech sample includes the characteristics induced by the secret (and distantly placed) recorder, room acoustics of the recording environment, and playback (i.e., loudspeaker) devices. Playback devices are nothing but the loudspeakers, which typically have the non-flat magnitude frequency response that is acting as bandpass filter [53]. The recording device affects similarly on the input signal.
 - b. In digital recording, the signal passes through an analog-to-digital converter (ADC), which uses the lowpass anti-aliasing filter. The spoofed

speech signal is processed through the anti-aliasing filter at least twice. It produces artifacts near the Nyquist frequency.

- c. Finally, room acoustics induces the reverberation effect, which causes temporal smearing due to the Short-Time Fourier Transform (STFT) [54–56].

1.3 Development of Standard Datasets for Voice Anti-Spoofing

Vulnerabilities and CMs for the other biometric modalities, such as face, fingerprint, etc. had been widely studied [47, 57–59]. The initiative for the voice biometric was taken by the research community and results in the organization of the first special session in *Spoofing and Countermeasures for ASV*, held during *INTERSPEECH 2013* [60]. At that opportunity, specific vulnerabilities and their CMs were presented in paper [61]. The search for a standard dataset, evaluation metrics, and protocols was also discussed in that paper, aiming at the development of an efficient CM system for unseen environments in the realistic scenarios [61]. This resulted in the development of the standard dataset, evaluation metrics, and protocols for the *ASVSpooF 2015 Challenge* - a special session during *INTERSPEECH 2015* [2]. It allowed for performance comparisons among various CM systems on a common evaluation platform across different sites. The SS- and VC-based spoofing attacks have been taken into consideration during this challenge. The second ASVSpooF challenge was organized during *INTERSPEECH 2017*, which motivated to develop CMs against replay attacks. During *INTERSPEECH 2019*, ASVSpooF 2019 challenge was organized, which focused on all the three major attacks, namely, SS, VC, and replay attacks. This edition of challenge consists of two scenarios, namely, Logical Access (LA), and Physical Access (PA). LA scenario includes the up-to-date SS- and VC-based systems (e.g., use of neural vocoders) to generate the spoof speech signals. PA scenario includes a controlled setup in the form of replay attacks, simulated using a range of real replay devices and carefully controlled acoustic conditions, which brings new insights into the replay spoofing problem. The latest edition of this challenge series, namely, ASVSpooF 2021 challenge includes DeepFake speech detection including modification in LA and PA scenarios of ASVSpooF 2019 challenge datasets. The replay configuration is modified as compared to the ASVSpooF 2019 challenge by introducing real and variable physical spaces. The ASVSpooF 2021 LA evaluation data

includes a collection of bonafide and spoofed utterances transmitted over a variety of telephony systems including Voice-over-IP (VoIP), and a public switched telephone network (PSTN), which utilizes the various speech *codecs*. Thus, the transmission channel and compression variability (due to codecs) in testing conditions is introduced in LA scenario.

These challenge corpora are designed to develop a CM for the ASV systems and not for VAs. A VA is said to be under replay attack, when an attacker tries to gain illicit access to the VA simply by replaying the voice commands of a genuine user. Hence, it is necessary to develop CMs for VAs as well. Although, the design of CM systems for ASV and VAs looks similar, there are important differences in anti-spoofing strategies for these systems. These differences are as follows [9]:

- In ASV applications, the close-distance (i.e., near-field speech) features can be preferred for SSD [9, 62]. On the other hand, these features are deteriorated due to longer distant (i.e., far-field) between actual speaker and microphone.
- VAs use microphone array for speech enhancement purpose, whereas ASV systems primarily use single microphone to record the speech samples. However, author believe that given the prevalence of VA systems, it is likely that they become a primary mean for ASV system including relevance of distant speech.
- Modern ASV systems follow strict ASV model, whereas VAs use less strict ASV model [63]. This makes it easier for the attacker to obtain a clean source recording to attack VAs.

To address this issue, *Realistic Replay Attack Microphone Array Speech Corpus* (ReMASC) and *Voice Spoofing Detection Corpus* (VSDC) datasets have been released, which are specifically designed to develop CMs against replay spoofing attack in VAs [9, 64].

1.4 Contributions of the Thesis

This thesis aims at developing effective CMs for SSD task, in particular, development of the handcrafted feature sets for the classification of genuine *vs.* spoof speech utterances. These CMs are built for anti-spoofing against SS, VC, and replay attacks. Furthermore, depending upon the distance between actual speaker and microphone, decision can be taken for appropriate feature extraction scheme.

For example, $CTECC_{max}$ should be utilized for long distance and pop noise detection algorithms should be utilized for short distance. These feature sets are either developed by applying the subband filtering on the speech signals or derived from the spectrogram representations.

1.4.1 Subband Filtering-Based Features

- **Enhanced Teager Energy-Based Cepstral Coefficients (ETECC)**

The proposed ETECC feature set is developed from Teager Energy Cepstral Coefficients (TECC) by compensating the *signal mass*. In Teager Energy Operator (TEO), we consider the approximation $\sin(\omega) \approx \omega$. The TECC feature set is extracted based on this approximation. However, the discriminative information for the replay SSD is prominently present in the mid- and high-frequency regions. Hence, ETECC feature set is proposed to accurately estimate the energies at higher frequencies, which is desirable for replay SSD task.

- **Cross-Teager Energy Cepstral Coefficients ($CTECC_{max}$)**

This feature set is proposed for a multichannel input using the concept of cross-Teager energy operator (CTEO), which measures the interaction between the two channels. To that effect, CTEO is used to estimate the most noisy transmission channels for effective SSD task. This idea develops $CTECC_{max}$ feature set for SSD task on VAs.

- **Energy Separation Algorithm-based Instantaneous Frequency estimation for Cochlear Cepstral Features (CFCCIF-ESA)**

Originally, Cochlear Filter Cepstral Coefficient Instantaneous Frequency (CFCCIF) feature set was proposed to develop CMs for SS and VC [65], where Instantaneous Frequencies (IFs) were estimated via analytic signal generation using Hilbert transform (HT). In our proposed feature set, IFs are estimated using Energy Separation Algorithm (ESA) to derive CFCCIF-ESA, which gives relatively better performance over CFCCIF.

1.4.2 Features Derived from Spectral Representations

- **Spectral Root Cepstral Coefficients (SRCC)**

In this work, deconvolution of a speech signal is performed by selecting an appropriate value of spectral root (i.e., γ) in the original homomorphic deconvolution framework [66]. In this technique, feature vectors are mapped

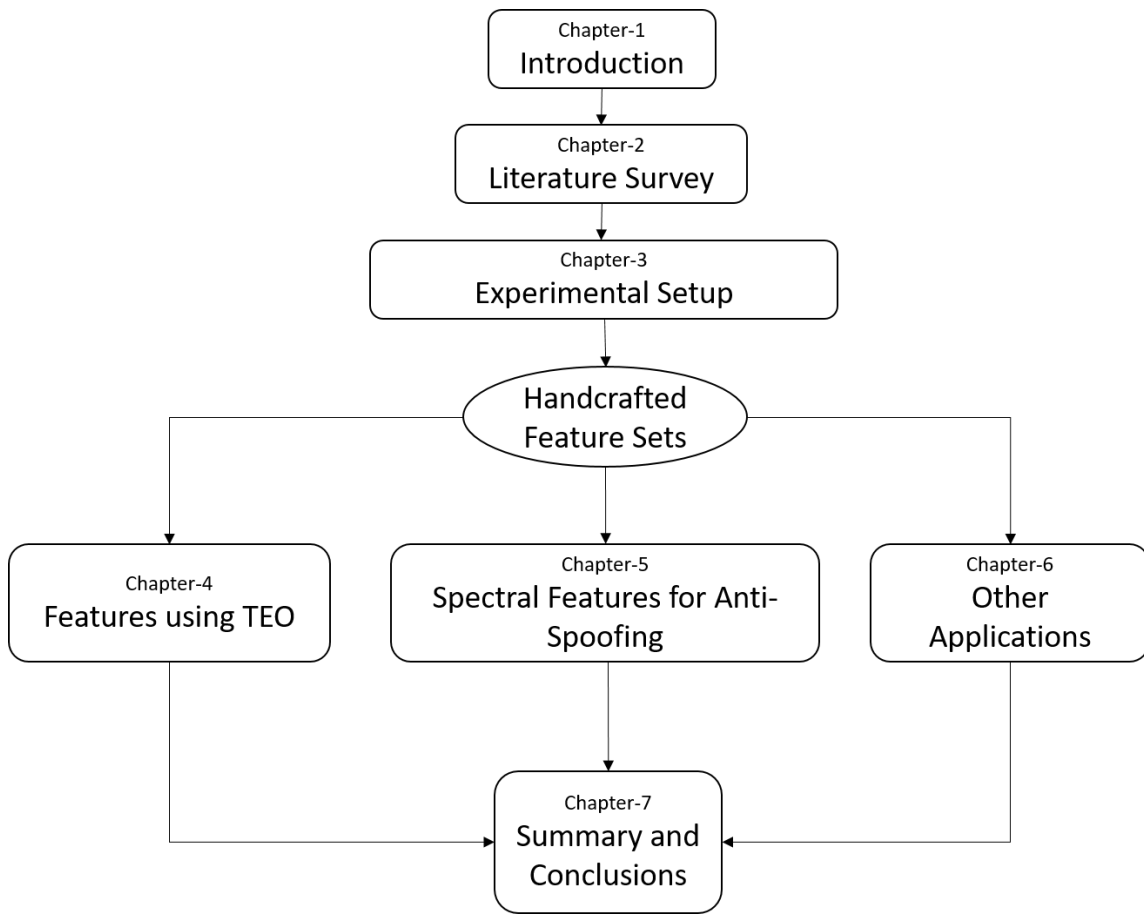


Figure 1.3: Flowchart Depicting Organization of this Thesis.

from one convolutional space to another convolutional space, where the convolved components are more easily separable than the previous space. This technique is used here for replay SSD task on ASV and VAs.

- **Constant-Q Transform (CQT) and Its Cepstral Representation**

In this work, CQT is used to extract the discriminative features from the pop noise regions in the speech utterance to detect liveness of the speaker. In realistic scenarios of speech production and perception, the perceived frequency of the speech signal doesn't have a constant frequency interval, rather it has a geometrical distribution [14]. Hence, we propose to use CQT as it mimics the speech perception mechanism. The constant Q-factor in CQT, yields better frequency resolution for lower frequencies and hence, it can effectively detect the pop noise, which posses the low frequency characteristics.

1.5 Organization of the Thesis

The organization of the rest of the chapters in the thesis is shown via a flowchart in Figure 1.3 and briefly described next.

Chapter 2 discusses the literature search of various methods proposed in the anti-spoofing literature to be able to position the contributions of this thesis work in the history of the problem.

Chapter 3 presents details of the experimental setup that is used to perform experiments reported in this thesis. These details include various datasets, feature extraction techniques, classifiers, performance evaluation metrics, and score-level fusion techniques. The variations in the data collection strategy along with the statistics is given in dataset Section. The theoretical description of state-of-the-art feature sets and classifiers is explained in corresponding Sections. Lastly, performance evaluation metrics and data fusion techniques are explained.

Chapter 4 discusses the performance of the various feature sets, which are derived using the concept of TEO. It includes ETECC, CTECC, and CFCCIF-ESA feature sets. Initially, basics of TEO are explained as it is the fundamental algorithm to derive these feature sets. Then, the development of each of these feature sets along with the explanation about their suitability to the intended SSD task is explained. Furthermore, the parameter tuning of the feature sets and classifiers is explained, which is followed by the experimental results.

Chapter 5 discusses the performance of the spectral-based feature sets, namely, CQT and SRCC feature sets, which are employed for replay SSD task. The CQT is utilized for Voice Liveness Detection (VLD), where the live speech is characterized by presence of the low-frequency pop noise. The capability of the CQT to capture the low frequency contents is exploited for locating the characteristics of pop noise. Furthermore, SRCC feature set employ *power-law* nonlinearity instead of the *logarithmic* nonlinearity, which may be more desirable for feature representation. The detailed explanation of the CQT and SRCC feature sets, their parameter tuning, and experimental setup is explained in Chapter 5.

Chapter 6 discusses the contributions of this thesis in the other speech technology applications. It includes the analysis of Cepstral Mean and Variance Normalization (CMVN) and various beamforming approaches for anti-spoofing, severity-level classification of the dysarthric speech, and classification of normal *vs.* pathological infant cry.

Chapter 7 presents the overall summary of the work presented in this thesis. This Chapter also discusses the applications, limitations of the present work, and

future research directions for the task of replay SSD on VAs and ASV systems.

1.6 Chapter Summary

In this chapter, we discussed the brief introduction of ASV field, its vulnerability towards possible spoofing attacks and hence, need to develop CM systems to alleviate such spoofing attacks. Furthermore, the initiative and efforts taken by research community to systematically pose the anti-spoofing problem was discussed. Finally, it gave the brief details of the contributions of this thesis, in particular, the development of the various feature sets to alleviate the issue of spoofing. The next chapter discusses in detail the development of the CM system, which includes various novel feature sets and classifiers on various datasets. Furthermore, research gap *w.r.t.* the literature and contributions of this thesis to fill this research gap is also discussed in the next chapter.

CHAPTER 2

Literature Search

2.1 Introduction

This chapter discusses the literature review on various widely used CM approaches on publicly available anti-spoofing datasets. ASV systems are prone to spoofing attacks based on VC [67], [68], SS [69], [70], impersonation [71], replay [72] [73], and twins [28]. In addition, the spoofing attacks on VAs have also gained a traction, and it leads to the development of the ReMASC and VSDC datasets with possible CM strategies for VAs. In subsequent Sections, the literature survey on the development of the various CM approaches *w.r.t.* various datasets are discussed. The literature search presented in this chapter will help us to understand various gap areas or research problems, which demands immediate attention from the community and thus, to position the key contributions made in this thesis as a humble step to fill this gap area.

2.2 Studies on SSD for SS and VC Attacks

In this Section, the performance of the various architectures (feature sets and classifiers) on the evaluation set of ASVSpooF 2015 challenge dataset is studied. This dataset is created using conventional vocoder generated SS- and VC-based spoof speech signals. It also includes the unit selection-based SS signals, which are difficult to identify as spoof signals. The evaluation set is chosen such that it should reflect the generalization capability in practical scenarios. The evaluation set consists of the *unknown attacks*, which successfully assesses the generalization capability of the SSD system. The more details of the dataset are explained in Chapter 3.2.1. Table 2.1 shows the results on the evaluation set for the various SSD architectures developed on ASVSpooF 2015 challenge dataset using conventional Gaussian Mixture Model (GMM)-based, Support Vector Machine (SVM), and deep neural network (DNN)-based classifiers.

The score-level fusion of CFCCIF and Mel Frequency Cepstral Coefficients (MFCC) feature sets was the best performing (winner) SSD system during ASVSpooof 2015 challenge organized during INTERSPEECH 2015 [29]. The CFCCIF feature set is an extension of Cochlear Filter Cepstral Coefficient (CFCC) feature set. The results of the individual feature sets, which are derived using cochlear filterbank, are also shown in Table 2.1. The *i*-vector feature set, which was originally developed for ASV task also produced relatively better results than the other feature sets [74,75]. In [76], several feature sets have been studied, such as MFCC, inverse-MFCC (IMFCC), Linear Frequency Cepstral Coefficients (LFCC), Rectangular Filter Cepstral Coefficients (RFCC), Linear Prediction Cepstral Coefficients (LPCC), Subband Spectral Flux Coefficients (SSFC), Spectral Centroid Magnitude Coefficients (SCMC), Subband Centroid Frequency Coefficients (SCFC), All-pole Group Delay Function (APGDF), and Relative Phase Shift (RPS). Among these feature sets, LFCC gave relatively better performance, which is shown in Table 2.1. The phase information being discriminative acoustic cue for the given SSD task, phase-based features, such as RPS and Modified Group Delay Cepstral Coefficients (MGDCC) were exploited by many participant teams in this challenge [76–79]. In another study, SVM with Generalized Linear Discriminant Kernel (GLDS-SVM) is utilized as a classifier, however, GMM shows the better performance than the GLDS-SVM classifier [80]. The study in [81] reports the development of the SSD system using various magnitude- and phase-based feature sets (referred to as M & P feats in Table 2.1, such as MFCC, product spectrum-based cepstral coefficients, MGDCC, weighted linear prediction group delay cepstral coefficients, linear prediction residual cepstral coefficients, Cosine Normalized Phase-based Cepstral Coefficients (CNPCC), and a combination of MFCC-CNPCC. Their best performing system obtained using score-level fusion gave 2.694 % Average Equal Error Rate (AEER). The analysis of the Linear Prediction (LP) error was also utilized for SSD of SS and VC-based spoof signals [82]. In [79], the various magnitude- and phase-spectrum-based feature sets, such as normalized unique local binary patterns (NULBP), modified group delay function features, and Cosine Normalized Phase Features (CNPf), are utilized to extract the complementary information via data fusion strategy for the SSD systems.

During the post evaluation of the ASVSpooof 2015 challenge, constant-Q cepstral coefficients (CQCC) feature set, which possesses a variable spectro-temporal resolution, gave the state-of-the-art results using $\Delta\Delta$ features [83]. Furthermore, the hierarchical scattering decomposition technique is proposed in [84], which derives the Scattering Cepstral Coefficients (SCC) feature set. It is viewed as a

generalization of all the constant-Q spectral decomposition, and produced the remarkable performance with reduced AEER of 0.18 %. In [85], performance of the proposed signal-based overlapped block transformation (SOBT), Mel warping overlapped block transformation (MOBT), signal-based frequency cepstral coefficients (SFCC), inverted signal-based frequency cepstral coefficients (ISFCC), and inverted Mel warping overlapped block transformation (IMOBT) feature sets are investigated for SSD task and consequently able to produce 0.86 % Equal Error Rate (EER) using ISFCC feature set. The excitation source-based features, namely, fundamental frequency (F_0) contour and strength of excitation (SoE) at the glottis, are also explored using GMM-based classification system, where score-level fusion is performed using MFCC and CFCCIF feature sets to produce better results [86, 87]. In [88], MFCC, modified relative phase (MRP) and MGDCC feature sets are utilized with GMM as a classifier. The score-level fusion of the SSD systems obtained using these three feature sets produced the 0.76 % AEER. In [89], three major types of artifacts related to magnitude, phase, and pitch (or fundamental frequency, F_0) variation introduced during the generation of synthetic speech, are exploited using three feature sets, namely, CQCC, APGDF, and fundamental frequency variation (FFV), respectively. These feature sets are concatenated, and acronymed as CAF. The novel FFV feature, introduced in [89] to extract pitch variation at the frame-level, provides complementary information to CQCC and APGDF.

In [90], two magnitude spectrum-based and five phase-based features have been applied using multilayer perceptron to train the SSD systems. The appropriate fusion of these SSD systems gave 2.62 % AEER on the evaluation set. In addition, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are employed, where CNN is used to extract the features from the speech signal, and RNN captures the long-term dependencies in those features for the SSD task [91]. The other feature sets utilized in the same study are: TEO critical band autocorrelation envelope, perceptual minimum variance distortionless response, and a more general spectrogram. In [92], log-magnitude spectrum and RPS feature sets are employed along with two DNN architectures, where one DNN architecture was used as a classifier, and the other used to extract bottleneck features. In [93], the feature sets derived from CQT, namely, short-term spectral statistics information (STSSI), Constant-Q Statistics-plus-Principal Information Coefficients (CQSPIC), octave-band principal information (OPI), and full-band principal information (FPI) are proposed, and these feature sets are trained using DNN classifier.

Table 2.1: Results (in % EER and % AEER) from the Literature for Various SSD Systems on ASVSpooF 2015 challenge Dataset. After [1].

System	Known Attacks						Unknown Attacks						All	
	S1	S2	S3	S4	S5	AEER	S6	S7	S8	S9	S10	AEER	AEER	AEER
CFCCIF + MFCC [29]	0.101	0.863	0.0	0.0	1.075	0.408	0.846	0.242	0.142	0.346	8.490	2.013	1.21	
i-vector [75]	0.004	0.002	0.0	0.0	0.013	0.008	0.019	0.0	0.015	0.004	19.57	3.992	1.96	
LFCC-DA [76]	0.027	0.408	0.0	0.0	0.114	0.110	0.149	0.011	0.074	0.027	8.185	1.670	0.89	
CQCC-A [83]	0.005	0.106	0.0	0.0	0.130	0.048	0.098	0.064	1.033	0.053	1.065	0.462	0.25	
CFCC [29]	0.04	1.39	0.00	0.00	2.30	0.75	1.04	0.12	0.06	0.21	12.28	2.74	1.74	
CFCCIF [29]	0.03	0.72	0.00	0.00	2.24	0.60	0.98	0.16	0.88	0.29	15.42	3.55	2.07	
CFCCIFS [29]	0.03	0.50	0.00	0.00	1.74	0.45	0.71	0.14	0.96	0.16	11.71	2.73	1.60	
ISOBT [85]	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000	0.000	16.840	3.374	1.687	
SFCC [85]	0.010	0.840	0.000	0.000	0.050	0.18	0.310	0.010	0.100	0.020	9.170	1.922	1.051	
ISFCC [85]	0.010	0.060	0.020	0.020	0.020	0.065	0.180	0.030	0.080	0.090	8.150	1.706	0.866	
FFV [89]	0.06	6.21	0.01	0.01	1.58	1.57	5.09	1.38	0.08	1.19	18.59	5.27	3.42	
CAF [89]	0.00	0.02	0.00	0.00	0.00	0.004	0.10	0.02	0.00	0.02	0.30	0.44	0.05	
MRP [88]	0.000	0.009	0.000	0.000	0.036	0.009	0.025	0.011	0.004	0.000	7.556	1.519	0.764	
SoE [86]	0.00	0.375	0.00	0.00	0.18	0.11	0.16	0.02	0.087	0.022	15.30	3.12	1.61	
SCC [84]	0.01	0.12	0.00	0.00	0.02	0.02	0.01	0.01	0.03	0.01	3.94	0.33	0.18	
RPS [77]	-	-	-	-	-	0.21	-	-	-	-	-	8.883	4.547	
RPS + MGDCC [78]	0.00	0.009	0.00	0.00	0.015	0.005	0.081	0.005	0.080	0.00	37.06	7.44	3.72	
M&P feats. * [81]	0.024	0.104	0.025	0.016	0.032	0.041	0.093	0.010	0.236	0.000	26.392	5.347	2.694	
M&P feats. * [79]	0.173	0.610	0.319	0.289	0.399	0.358	0.906	0.242	0.417	0.246	28.581	6.078	3.218	
s-vector [94]	-	-	-	-	-	0.058	-	-	-	-	-	4.998	2.528	
M & P feats. [90]	0.00	0.00	0.00	0.00	0.01	0.002	0.01	0.00	0.00	0.00	26.1	5.22	2.62	
Spectro/CNN [91]-a	0.08	0.19	0.02	0.03	1.26	0.31	1.48	0.68	0.01	0.16	26.83	5.83	3.07	
Spectro/RNN [91]-b	1.21	0.79	0.24	0.39	1.77	0.87	0.87	0.96	0.04	0.41	17.97	4.05	2.46	
Spectro/CNN + RNN [91]-c	0.16	0.50	0.03	0.03	1.38	0.40	0.85	0.91	0.03	0.59	14.27	3.33	1.86	
Fusion [91]-(a+b+c)	0.09	0.29	0.00	0.00	0.99	0.27	0.64	0.71	0.00	0.29	11.67	2.66	1.47	
(Spectrum + RPS)/ DNN [92]	0.021	0.031	0.021	0.023	0.031	0.025	0.038	0.032	0.041	0.021	40.708	8.168	4.096	
CQSPIC-A [93]	0.00	0.00	0.00	0.00	0.004	0.00	0.00	0.00	0.008	0.00	0.368	0.075	0.038	
CQUEST-DA [95]	0.00	0.00	0.00	0.00	0.009	0.00	0.005	0.004	0.089	0.00	0.456	0.110	0.056	
CMC-A [96]	0.00	0.00	0.00	0.00	0.004	0.00	0.005	0.00	0.026	0.00	0.221	0.050	0.026	

* M&P feats. is acronymed for fusion of the various magnitude- and phase-based features.

Among these features, CQSPIC-A shows the remarkable performance as shown in Table 2.1 [93], where, -A refers to $\Delta\Delta$ -features. Other study in [95] proposes a subband transform rather than the fullband transform on CQT with three different scales, i.e., linear, octave, and Mel scale to derive three feature sets, namely, Constant-Q Equal Subband Transform (CQ-EST), Constant-Q Octave Sub-band Transform (CQ-OST), and Discrete Fourier Mel Subband Transform (DF-MST). The CQ-EST-DA (-DA refers to combination of Δ and $\Delta\Delta$ features) feature set with DNN classifier gave the 0.056 % AEER. Recently, the Multilevel Transform (MLT) is applied to CQT to derive Constant-Q Multi-level Coefficients (CMC) [96]. It can be observed from Table 2.1 that the CQCC and feature sets derived from the CQT are producing relatively better results and thus, it shows their generalization capability to perform better on unseen spoofing attacks.

2.3 Studies on Real Replay SSD

This Section discusses various architectures proposed on ASVSpooof 2017 challenge dataset, which comprise real replay scenario. The spoof speech utterances in this dataset are collected from seven environments, namely, anechoic room, analog wire, balcony, canteen, home, office and studio. The recordings in various environments produces the variability in the spoof speech signals in terms of environments [3]. Furthermore, various recording/playback devices also employed to increase the variability. The details of the dataset are discussed in Chapter 3. Some notable contributions on ASVSpooof 2017 database are as shown in the Table 2.2. The best possible performance is achieved by MGDCC feature set with Residual Networks (ResNet) as a classifier [97]. Siamese embeddings are also employed for replay spoof detection in [98]. In [99], Mel Filterbank Slope (MFS) and Linear Filterbank Slope (LFS) feature sets are proposed, which captures low frequency information corresponding to that of the low quality recording devices and high frequency information corresponding to that of a high quality recording device, respectively. Furthermore, DenseNet-long short-term memory (LSTM) is proposed for replay anti-spoofing in [100]. Spectral envelope centroid frequency and spectral envelope centroid magnitude features are introduced in [101], which are based on *Spatial Differentiation*. The novel Adaptive Relative Phase (ARP) and Adaptive Frequency Cepstral Coefficients (AFCC) feature sets are proposed in [102] along with the design of the attention-based adaptive filters. Furthermore, auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM) with the pre-emphasized speech signals is exploited to extract

Amplitude Modulation and Frequency Modulation (AM and FM)-based features. Cepstral processing is performed on ConvRBM-based short-time AM and FM features to give AM-ConvRBM-CC and FM-ConvRBM-CC feature sets [103]. Other feature sets and classifiers, which are used for ASVSpooof 2015 challenge dataset can be observed in Table 2.2 and details regarding the architectures can be studied from corresponding references [17, 104–107].

2.4 Studies on SSD for SS, VC, and Simulated Replay

This Section discusses various architectures proposed on ASVSpooof 2019 challenge dataset, which comprises the three major spoofing attacks, namely, SS, VC, and simulated replay attacks. The brief description of the dataset can be studied in Chapter 3. The simulated (controlled) replay and Neural Network (NN)-based SS/VC-generated spoofing signals are utilized in this challenge. In [108], various systems are combined, which are based on unified/entire feature maps along with variations of squeeze-excitation and residual networks, such as SENet34, SENet50, ResNet, Dialated Resnets, and attention filtering network. *i*-vectors extracted from MFCC, IMFCC, CQCC, sub-band centroid magnitude coefficient (SCMC) features are utilized with CNN, RNN, Wave-U-Net, GMM, and SVM classifiers [109]. The DKU system in [110] presents the CM system in PA scenario through data augmentation, which utilizes CQCC, LFCC, IMFCC, STFT-gram, Group Delay-(GD) gram, joint-gram as feature sets, and ResNet as a classifier. Bayesian neural networks are employed in [111] considering their capability to generalize the model. Angular margin-based softmax activation was utilized in Light-CNN (LCNN) along with CQT, LFCC, Fast Fourier Transform (FFT), and Discrete Cosine Transform (DCT) representations [112]. In another approach, Log-CQT, Log Mel Spectrogram (LMS), Phase features, *i*-vector, and Variational Autoencoder (VAE)-log-CQT representations are utilized with ResNet, LCNN with multilabel output, and context gate CNN (CGCNN). In [113], Single Frequency Cepstral Coefficients (SFCC), Zero-Time Windowing Cepstral Coefficients (ZTWCC), and Instantaneous Frequency Cepstral Coefficients (IFCC) feature sets are utilized with the conventional GMM-based classifier. CQT and its variants are exploited in [114, 115]. Utterance-level embeddings are extracted using a Light Convolutional Gated Recurrent Neural Network (LC-GRNN) [116]. In another study, combination of the VGG and LCNN architecture is exploited along with MFCC, CQT, CQCC, and power spectrogram representations [117]. The results obtained with above mentioned approaches are shown in Table 2.3.

Table 2.2: Results Obtained on ASVSpooof 2017 Version 2.0 Dataset for Various Systems Reported in the Literature.

Authors (Source)	Feature Sets	Classifiers	% EER	
			Dev	Eval
<i>Tom et. al.</i> [97]	MGDCC	ResNet	0	0
<i>Sriskandaraja et. al.</i> [98]	Siamese Embedding Features	GMM	-	6.40
<i>Lawrentyeva et. al.</i> [17]	Spectrogram	CNN and RNN	3.95	6.73
	Spectrogram	LCNN	4.53	7.37
<i>Saranya et. al.</i> [99]	MFS	GMM	3.58	7.82
	LFS	GMM	5.13	9.82
<i>Weicheng Cai et al.</i> [104]	CQCC (Baseline)	GMM	10.25	22.39
<i>Patil et. al.</i> [105]	MFCC	GMM	26.78	26.31
	CFCCIF	GMM	12.98	14.77
	LFCC	GMM	16.76	13.9
<i>Kamble et. al.</i> [106]	TECC	GMM	9.55	11.73
	Hybrid Feature	GMM	8.67	25.63
<i>Lian Huang et. al.</i> [100]	Hybrid Feature	DenseNet	5.62	12.39
	Hybrid Feature	LSTM	9.45	15.64
	CQCC	DenseNet	7.65	17.73
	MFCC	DenseNet	6.77	15.86
	CQCC	DenseNet-LSTM	3.87	12.64
<i>Buddhi Wickramasinghe et. al.</i> [101]	CF	GMM	-	10.84
	CM	GMM	-	10.93
<i>Meng Liu et. al.</i> [102]	AFCC	GMM	4.01	27.80
	ARP	GMM	9.11	12.65
	CQCC+AFCC	-	3.57	28.02
	CQCC+ARP	-	2.26	12.58
	AFCC+ARP	-	2.23	11.95
	ARP+AFCC+CQCC	-	2.20	11.43
<i>Roberto Font et al.</i> [107]	RFCC	GMM	6.91	11.90
	LPCC	GMM	5.94	25.20
	SCFC	GMM	24.51	24.83
	SCMC	GMM	9.32	11.49
	SSFC	GMM	12.81	22.38
<i>Sailor et. al.</i> [103]	AM-ConvRBM-CC (S1)	GMM	2.92	12.76
	FM-ConvRBM-CC (S2)	GMM	5.44	14.96
	S1 + S2	GMM	0.82	8.89

Table 2.3: Results Obtained on ASVSpooof 2019 Dataset (LA and PA Scenario) for the Various Architectures in the Literature.

Source	Feature Sets	Classifiers	LA				PA			
			Dev		Eval		Dev		Eval	
			% EER	t-DCF	% EER	t-DCF	% EER	t-DCF	% EER	t-DCF
[108]	Unified/Whole Feature Map	SENet34, SENet50, ResNet, Dialated ResNets	0	0	6.70	0.155	0.129	0.003	0.59	0.016
[109]	<i>i</i> -vectors extracted from MFCC, IMFCC, CQCC, SMCC	CNN, CRNN, GMM, SVM Wave-U-Net	0	0	2.64	0.0755	4.85	0.1316	5.43	0.1465
[110]	CQCC, LFCC, STFT-gram IMFCC, GD- & joint-gram	ResNet	-	-	-	-	0.24	0.0064	0.66	0.0168
[111]	Mel Spectrogram	Baysian NN (CNN, LCNN)	-	-	-	-	0.78	0.0170	0.88	0.0219
[112]	CQT, LFCC, FFT, DCT	LCNN	0	0	1.84	0.0510	0.0154	0.0001	0.54	0.0122
[130]	CQT, Phase Vector LMS, <i>i</i> -vector	ResNet, LCNN, CGCNN	0.90	0.027	3.56	0.1118	0.0049	0.16	1.16	0.03550
[113]	SFCC, ZTWCC, IFCC	GMM	0	0	0.1239	4.92	0.2169	10.11	0.2810	12.20
[115]	CQCC, eCQCC, ICQCC, CMC, CQ-EST, CQ-OST, CQSPIC	GMM, DNN	0	0	4.13	0.1264	1.26	0.027	5.95	0.1381
[116]	Embedding Extracted using LC-GRNN	SVM, LDA, PLDA	0	0	6.28	0.1523	0.73	0.0203	2.23	0.0614
[117]	MFCC, CQT, CQCC Power Spectrogram	VGG, LCNN	0	0	8.01	0.2080	0.66	0.0170	0.0372	1.51

Furthermore, Butterfly Unit (BU) for multitask learning is employed in [118]. *x*-vector embeddings extracted from MFCCs with CNN classifier are utilized in [119]. Other possible approaches for ASVSpooof 2019 dataset can be studied from [4, 120–129].

2.5 Studies on SSD for SS, VC, Replay, and DeepFake Attacks

This is the 4th edition of the bi-annual ASVSpooof challenge campaign, and it also includes the DeepFake speech detection task along with LA and PA scenarios in ASVSpooof 2019 challenge. However, it introduces the difficulty by introducing transmission channel and compression variability in LA scenario, and real physical spaces in PA scenario. The further details of the dataset can be studied from Chapter 3. In [131] spectro-temporal Graph Attention Network (GAT), which learns the relationship between acoustic cues spanning different sub-bands and temporal intervals is proposed. The model-level graph fusion of spectral (S) and temporal (T) sub-graphs and a graph pooling strategy (RawGAT-ST) is employed to achieve the better performance. SELCNN network is proposed in [132], which inserts squeeze-and-excitation (SE) blocks into a LCNN to enhance the capacity of hidden feature selection. Then, multitask learning (MTL) frameworks is implemented using SELCNN followed by the bidirectional long short-term memory (Bi-LSTM) as the basic model. Other evidences of squeeze excitation networks can be studied in [133, 134]. In [135], the difference between the vocoder-filtered audio

and the original audio is used as the input feature and multiple outlier detection model is adopted as the backend classifier. In the other study, raw differentiable architecture search system is employed for DeepFake and anti-spoofing [136]. Various data augmentation techniques over CQT are utilized in [137] for DeepFake and anti-spoofing. Multiple-point input for CNNs is explored in [138]. The interesting study, which analyzes the significance of the leading and trailing silence regions, can be studied in [139]. In the other studies, Time Delay Neural Networks (TDNN) and its variants are utilized for anti-spoofing [140, 141]. Furthermore, the activation functions are analyzed for robust end-to-end spoofing attack detection system in [142]. A Higher-Order Statistics Pooling (HOSP) method for extracting the utterance-level embedding is adopted in [143]. Other studies on ASVSpooF 2021 dataset can be studied in [144, 145].

2.6 Other Datasets Used

Other than the datasets released as a part of ASVSpooF challenge campaigns, research community also contributed in designing the other datasets for anti-spoofing research, such as POp noise COrpus (POCO), ReMASC, and VSDC. In particular, POCO dataset is designed for Voice Liveness Detection (VLD), whereas ReMASC and VSDC are designed for replay SSD task on VAs. This Section is dedicated to the various CM approaches developed on these datasets.

Pop noise can be utilized as acoustic signature of the live (or genuine) speaker. To that effect, a few attempts are reported to utilize the pop noise as a signature to detect the presence of the live speaker in front of the ASV system. In [34], authors collected their own dataset and proposed two STFT-based approaches for the VLD task. This study was extended in [62] with more detailed experimental setup and analysis of results. In one of the approach, features are extracted from the low frequency regions because pop noise has the low frequency characteristics. In the second approach, subtraction is performed on the STFT spectrogram obtained from pop noise and corresponding non-pop noise utterance. The study is extended in [146, 147] by selecting the pop noise-specific phonemes in the utterances, and it proved to be the efficient approach for pop noise detection. Furthermore, in [148], a robust software, namely, VoicePop is designed and implemented on smartphones for anti-spoofing, where Gammatone Frequency Cepstral Coefficients (GFCC) are utilized for feature extraction. During INTERSPEECH 2020, POCO dataset was released along with the baseline algorithm, which is specifically designed for pop noise detection [8]. Using this dataset, a few architectures

on VLD can be studied [35, 149–152], which utilizes Modified Group Delay Functions (MGDF), CQT, STFT as feature representations along with GMM, SVM, and CNN as classifiers.

To address the replay SSD for VAs, ReMASC and VSDC datasets are developed [9, 153]. Both the datasets consist of microphone array for distant (far-field) speech recognition. During design of the VSDC dataset, single- and multi-order replay was also considered using drop-in feature in a VA, which will be useful to design CM in that context [153]. The CQCC-GMM baseline is also provided along with the dataset. In [154], cross-dataset performance is assessed for one-point and two-point replay SSD task. One of the architecture developed on ReMASC dataset can be studied in [12]. In addition, this thesis contributes in assessing the performance of the proposed ETECC, CTECC, and SRCC feature sets on the ReMASC dataset.

2.7 Gap Area in the Anti-Spoofing Literature

Numerous gap research areas can be found via the literature search presented in the earlier Sections of this Chapter. However, this thesis picks up the following research gaps and attempts to alleviate those gaps:

- In the earlier studies, TECC feature set was proposed for SSD task [41]. In TECC, TEO estimates the instantaneous (or running estimate of) energy of the signal by considering the approximation $\sin(\omega) \approx \omega$, which is applicable (or valid) for the lower frequencies. Whereas, the distortions introduced due to replay mechanism is *bandpass* in nature, i.e., high frequency regions are also affected in replay spoof signals. These distortions in high frequency regions could not captured using TECC.
- The proposed CFCCIF feature set combines magnitude and phase information and produces an efficient representation for building the CM against VC- and SS-based spoofing attacks. Phase information is presented in the form of Instantaneous Frequency (IF), which is extracted using HT. However, it could not estimate the IF instantaneously and requires the entire segment of the speech signal.
- The earlier proposed anti-spoofing approaches were designed for ASV system. Whereas, the utility of the voice biometrics in VAs is increasing exponentially in recent times. However, less attention is being given to develop the CMs for anti-spoofing on VAs.

- Intuitively, anti-spoofing can be implemented by emphasizing the characteristics of the live speaker. The presence of the pop noise was utilized as characteristics of the live speaker. The pop noise has the low frequency characteristics, which should be effectively captured for the VLD task. In the literature, this pop noise is tried to emphasize using STFT spectrogram, whose resolution at low frequency region is moderate due to linear spacing of the frequency bins.
- In MFCC, *logarithmic cepstrum* is utilized, which cannot be tuned based on the system property of the speech signal. Hence, it may sometime fail to capture the system property effectively using a lesser number of cepstral coefficients.

2.8 Key Contributions in the Thesis

Considering the gap areas mentioned above and discussions *w.r.t.* contributions in this thesis in Chapter 1, a humble attempt is made to fill these gaps in order to improve the performance of the CMs against spoofing attacks on ASV and VAs. To that effect, the following feature sets are proposed in this thesis:

- ETECC feature set is proposed, which is derived using the novel energy operator, namely, Enhanced Teager Energy Operator (ETEO). It uses recently introduced concept of the *signal mass* and estimates the energies more accurately for the entire frequency range, more so, for high frequency regions.
- In proposed CFCCIF-ESA feature set, IFs are estimated using ESA. This approach of IF estimation helps to estimate the IFs more accurately and instantaneously. The significant improvement is observed for CFCCIF-ESA feature set over the existing CFCCIF feature set.
- The acoustic information in the microphone array is exploited using CTECC_{max} feature set for replay SSD task on VAs. In the proposed CTECC_{max} framework for anti-spoofing, the cross-Teager energy among the subband channels of microphone array is maximized to extract the noise distortions introduced because of the acoustic environment.
- The low frequency characteristics of the pop noise is emphasized by CQT. In particular, the parameters of the CQT are tuned to obtain the desired frequency resolution in low frequency regions. The geometrical spacing be-

tween the frequency bins of CQT helps to achieve the high frequency resolution in the low frequency regions.

- SRCC feature set is proposed to effectively estimate the system information the feature set with a less number of coefficients.

2.9 Chapter Summary

This Chapter discusses the literature search on voice anti-spoofing for ASV and VAs. It generally includes the development of the CM systems against spoofing attacks. The first profound step was taken during the first special session in *Spoofing and Countermeasures for ASV*, held during *INTERSPEECH 2013*. After that, four bi-annual ASVSpooF challenges were organized, which provided the common platform to validate the performance of the CM systems. This chapter provided the overview of the various approaches proposed on those datasets including the other datasets, such as BTAS 2016, POCO, and ReMASC datasets. Given this literature search, few research gaps were observed followed by the contribution in the light of these gap areas. The next chapter discusses the details of experimental setup used in this thesis work.

CHAPTER 3

Experimental Setup

3.1 Introduction

This chapter discusses various components of experimental setup that are extensively used in various experiments reported in this thesis. In particular, various datasets, feature sets, classifiers, evaluation metrics, and score-level fusion strategies are briefly discussed in this chapter. Various datasets, such as ASVSpooof 2015, ASVSpooof 2017, ASVSpooof 2019, POCO, and ReMASC datasets had been utilized in order to investigate significance of the proposed handcrafted feature sets. Furthermore, CQCC, LFCC, and MFCC are explained in brief as these are the baseline feature sets in ASVSpooof challenge campaigns during INTERSPEECH conferences. In addition, TECC, Squared Energy Cepstral Coefficients (SECC), and cepstrals have been discussed as these feature sets were utilized for comparison with the proposed feature sets. In addition, various classifiers, such as GMM, SVM, CNN, LCNN, and ResNet are discussed. Finally, the score-level (data) fusion strategies are also discussed.

3.2 Standard Anti-Spoofing Corpora

As discussed in Chapter 2, significant efforts have been made by the ASVSpooof challenge organizers to release standalone and statistically meaningful corpora for anti-spoofing research. Such corpora are very important *w.r.t.* reproducible research - a global concern. Until 2013, there were no standard datasets, methodology, or evaluation methods for anti-spoofing research and hence, this was a serious hindrance to design an effective and consistent CM system *w.r.t.* reproducible research. Reproducibility is a recent concern expressed by computational researchers working in the signal processing and machine learning field [155]. In this context, the study reported in [156], inspired the special session, "Reproducible Signal Processing Research", at the *IEEE International Conference on Acous-*

tics, Speech, and Signal Processing (ICASSP) 2007. A study from that session by Kovacevic [157] and a recent study in [158] revealed the reproducibility crisis and necessary initiative towards open codes and data. In particular, research is reproducible if all the necessary information that is related to the work, including, but not limited to, text, data, and code is made available so that anyone, anywhere can reproduce the results at any point of time given similar computational resources [159].

The speaker recognition, one of the potential research issue in speech signal processing, continues to be data-driven field and so is the anti-spoofing field. In particular, results obtained for ASV or CM/SSD systems are meaningless if recording conditions of the corpus are not known. To that effect, the first significant step taken towards the development of CM systems was during the first special session in *Spoofing and Countermeasures for ASV*, held during *INTERSPEECH 2013* [60]. One of the studies in that session discussed the need for a standard dataset, evaluation metrics, and protocols aiming at the development of an efficient CM system for unseen environments in realistic scenarios [61]. This resulted in the development of the ASVSpooF challenge campaigns. In particular, ASVSpooF 2015 challenge considered the SS- and VC-based spoofing attacks, whereas ASVSpooF 2017 challenge considers the replay attack. Furthermore, ASVSpooF 2019 challenge considers all the three kinds of attacks in two scenarios, namely, LA and PA. The LA-scenario includes state-of-the-art SS- and VC-based approaches during those time, and the PA-scenario consists of simulated replay attack. ASVSpooF 2019 challenge includes the DeepFake challenge in addition to the LA and PA scenarios in ASVSpooF 2019 dataset. Here, we first discuss details of various corpora used in this thesis.

3.2.1 ASVSpooF 2015 Challenge Dataset

This dataset was developed for ASVSpooF 2015 Challenge campaigns, which is organized during *INTERSPEECH-2015* to address the SS- and VC-based spoofing attacks.¹ The genuine speech samples are recorded from a total of 106 subjects, which consists of 45 male and 61 female speakers. Genuine utterances are collected with minimum background and transmission channel noise. Ten different algorithms of SS and VC (details given in Table 3.2) are utilized to generate spoofed speeches [2, 160].

These spoofed signal classes are labelled as S_1, S_2, \dots, S_{10} . Among these, train-

¹The ASVSpooF 2015 dataset along with the standard protocols is publicly available at <https://datashare.ed.ac.uk/handle/10283/853> [Last Accessed: June 1, 2022].

Table 3.1: Statistics of the ASVSpooof 2015 Challenge Dataset Partition. After [2].

Subset	# Speakers		# Utterances		Spoofing Approaches
	M	F	Genuine	Spoof	
Train	10	15	3750	12625	S1 to S5
Dev	15	20	3497	49875	S1 to S5
Eval	20	26	9404	184000	S1 to S10

Dev = Development, Eval = Evaluation.

Table 3.2: Spoofing Algorithms Implemented in the ASVSpooof 2015 Challenge Dataset. After [2].

Subset	# Utterances			Vocoder	Spoofing Algorithm
	Train	Dev	Eval		
Genuine	3750	3497	9404	None	None
S1	2525	9975	18400	STRAIGHT	VC
S2	2525	9975	18400	STRAIGHT	VC
S3	2525	9975	18400	STRAIGHT	SS
S4	2525	9975	18400	STRAIGHT	SS
S5	2525	9975	18400	MLSA	VC
S6	0	0	18400	STRAIGHT	VC
S7	0	0	18400	STRAIGHT	VC
S8	0	0	18400	STRAIGHT	VC
S9	0	0	18400	STRAIGHT	VC
S10	0	0	18400	None	SS

ing and development (Dev) subsets uses S1 to S5 algorithms, whereas evaluation (Eval) subset includes S1 to S10. As the spoofing algorithms in the Dev set are seen by the training model of SSD system, these spoofing attacks are called as *known* attacks. Whereas the spoofing algorithms not seen by the training model in the Eval set are known as *unknown* attacks. Among these spoofing algorithms, S3, S4, and S10 uses speech synthesis algorithms, and the others are VC-based approaches. S1-S9 uses vocoder-based algorithms, and S10 uses unit selection-based approach for speech synthesis. Two vocoders, namely, STRAIGHT [161] MLSA [162, 163], are utilized to implement vocoder-based algorithm. The details of these spoofing algorithms can be studied in [2]. The statistics of the partition of ASVSpooof 2015 dataset into training, Dev, and Eval set is provided in Table 3.1. Furthermore, the statistics of the partition of the dataset *w.r.t.* spoofing algorithms is shown in Table 3.2.

Table 3.3: Statistics of the ASVSpooF 2017 Dataset for the Environment-Independent Case. After [3].

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

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

Table 3.4: Distribution of Spoof Speech Utterances Among the Environments in ASVSpooF 2017 Dataset

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

- : Not applicable

3.2.2 ASVSpooF 2017 Challenge Dataset

This dataset was released for ASVSpooF 2017 challenge organized during INTER-SPEECH 2017, and later it was made publicly available [164]. However, data anomalies, such as silence regions and zero values (i.e., artificial silence regions), have been noticed by the challenge organizers and then these anomalies are fixed in the second version of the dataset, which is known as ASVSpooF 2017 Version 2.0 dataset [3]². In this dataset, genuine utterances are selected from the RedDots corpus, which is designed for text-dependent ASV using *ten* prompt sentences [165]. Replay spoof signals are generated in 177 sessions using various acoustic environments and heterogeneous devices. The standard partition of the dataset into training, Dev, and Eval is done as shown in Table 3.3 [3]. The dataset comprised of 61 distinct replay configurations, which are a combination of a playback device, a recording device, and an acoustic environment.

These configurations pose varying amount of difficulty to detect the replay spoof speech signal and hence, poses the varying threat depending upon the replay configurations. ASVSpooF 2017 dataset is collected in a total of 26 different

²The ASVSpooF 2017 V2.0 dataset along with protocols and extended metadata is available online at <https://datashare.ed.ac.uk/handle/10283/3055> {Last Accessed: June 1, 2022}.

acoustic environments. The two *balcony* and one *canteen* are high ambient noise environments, which are relatively easy for SSD and hence, it produces low threat to the ASV system. The eight *home* and *ten* office conditions consists medium ambient noise-level and hence, produce medium-level threat to the ASV system. However, an anechoic room, studio, and analog wire recordings exhibit very low additive noise and hence, poses the graver threat to ASV system. The distribution of the spoof utterances among the various environments is shown in Table 3.4. Furthermore, 26 playback devices with varying quality are utilized. These playback devices may be consumer grade replay devices with smaller loudspeakers (i.e., low threat), consumer grade replay devices with larger loudspeakers (i.e., medium threat), and professional audio equipments (i.e., graver threat). Replay samples are recorded with 25 various recording devices, and categorized into low, medium, and graver threats, depending upon the quality of recording device. In addition, it can be observed that the replay environments, which poses the graver threats, are included in the Eval set so that the CM models can be assessed on difficult unknown attacks.

The experiments are also performed for environment-dependent scenario on ASVSpooF 2017 dataset, where target environment is seen by the defense model. In this case, training and testing are performed on each individual environment. The distribution of the number of spoof speech utterances for each environment is varying and shown in Table 3.4. To develop an individual environment-dependent SSD system, half of the spoofed speech utterances for the corresponding environment were chosen for training, whereas the remaining half were used for testing the model performance. To train the genuine and the spoofed speech models, an equal number of utterances were selected.

3.2.3 ASVSpooF 2019 Challenge Dataset

ASVSpooF 2019 challenge considers all the three major spoofing attacks, namely, SS, VC, and replay spoofing attacks [4]. However, this challenge is explored in two scenarios, namely, PA and LA. PA addresses the replay spoofing attacks, whereas LA addresses SS- and VC-based spoofing attacks³. Though ASVSpooF 2019 dataset considers a similar kind of spoofing attacks as that of ASVSpooF 2015 and ASVSpooF 2017 challenge campaigns, there are important modifications performed in 2019 edition than the previous challenges. In ASVSpooF 2015 challenge, traditional vocoders, such as STRAIGHT and MLSA, were utilized for the gener-

³The ASVSpooF 2019 dataset, along with its protocols, is available at <https://datashare.ed.ac.uk/handle/10283/3336> {Last Accessed: June 1, 2022}.

Table 3.5: Statistics of the ASVSpooof 2019 Dataset. After [4].

# Subset	# Speakers		# Utterances			
	# Male	# Female	Logical Access (LA)		Physical Access (PA)	
			# Bonafide	# Spooof	# Bonafide	# Spooof
Train	8	12	2580	22800	5400	48600
Dev	8 (4 target, 4 non-target)	12 (6 target, 6 non-target)	2548	22296	5400	24300
Eval	30 (21 target, 9 non-target)	37 (27 target, 10 non-target)	7355	63882	18090	116640

ation of SS and VC spooofs. Whereas in ASVSpooof 2019 challenge dataset, neural network-based vocoders are utilized, which produces (or synthesizes) the voice comparable to that produced by the humans. In ASVSpooof 2017 challenge, the speech signal recordings are of real replayed spooofing attacks. The use of uncontrolled setup in ASVSpooof 2017 challenge made it difficult to analyze the results. Whereas, ASVSpooof 2019 PA dataset consists of replay attacks simulated using a range of real replay devices and carefully controlled acoustic conditions. This controlled setup brings new insights into the replay spooofing problem *w.r.t.* possible analysis of performance of SSD system for real *vs.* simulated replay. ASVSpooof 2015 and ASVSpooof 2017 editions of ASVSpooof challenge focused on the development and assessment of standalone CMs, whereas ASVspooof 2019 challenge adopted for the first time a new ASV-centric performance metric in the form of the tandem Detection Cost Function (t-DCF).

The ASVSpooof 2019 dataset is adapted using Voice Cloning Toolkit (VCTK) corpus [166], which is recorded in a hemi-anechoic chamber at a sampling rate of 96 kHz. The speech signals in VCTK corpus are downsampled to 16 kHz with a resolution of 16 bits-per sample, and considered as genuine speech signals in ASVSpooof 2019 dataset. The speakers are partitioned into the training, Dev, and Eval sets in ASVSpooof 2019 dataset as shown in Table 3.5. Furthermore, the Multi-speaker Multi-style Voice Cloning Challenge (M2VoC) was organized during ICASSP 2021, which provides a common sizable dataset as well as a fair testbed for the benchmarking of the popular voice cloning task [167]. The goal of this challenge is to adapt an average Text-to-Speech (TTS) model to the stylistic target voice with limited data from the target speaker, evaluated by speaker identity and speaking style similarity. The challenge consists of two tracks, namely, few-shot track and one-shot track, where the participants are required to clone multiple target voices with 100 and 5 samples, respectively.

Table 3.6: Algorithms for LA Spoofing Systems. Here, * Indicates Neural Network-Based Algorithm. After [4].

Algorithm	Input	Waveform Generator
A01	Text	WaveNet* [168]
A02	Text	WORLD [169]
A03	Text	WORLD
A04	Text	Waveform Concat.
A05	Speech (human)	WORLD
A06	Speech (human)	Spectral Filtering + OLA
A07	Text	WORLD
A08	Text	Neural Source-Filter*
A09	Text	Vocaine
A10	Text	WaveRNN*
A11	Text	Griffin-Lim
A12	Text	WaveNet*
A13	Speech (TTS)	Waveform Filtering
A14	Speech (TTS)	STRAIGHT
A15	Speech (TTS)	WaveNet*
A16	Text	Waveform Concat.
A17	Speech (human)	Waveform Filtering
A18	Speech (human)	MFCC Vocoder
A19	Speech (human)	Spectral Filtering + OLA

3.2.3.1 Logical Access (LA)

In this scenario, fraudulent person tries to attack by gaining the access of the ASV system. For example, in telephone banking, an attacker may acquire the access and try to breach the ASV using the digital copy of the voice of the authentic speaker. It includes the SS- and VC-based attacks, which can be performed post-sensor. The SS and VC technologies are exploited to transform the given text and voice, respectively, to the target speaker. The statistics of the ASVSpooF 2019 LA dataset is shown in Table 3.5 [4]. The spoof utterances in training and Dev sets are generated using four TTS and two VC algorithms. Here, spoofing utterances in the Dev set uses the similar algorithms as that of the training set. Hence, Dev set comprises known attacks. Spoofed samples in Eval set are generated using 7 TTS and 6 VC algorithms, where 2 algorithms are utilized for known attacks, and 11 algorithms are utilized for unknown attacks. In total, 19 TTS and VC algorithms are utilized, denoted as A01 to A19, are described in brief as shown in Table 3.6.

Table 3.7: Parameter Settings for Acoustic Configurations to Generate Simulated Replay Spoofs ASVSpooF 2019 Challenge Dataset. After [5].

	a	b	c
Room Size (in m^2)	2-5	5-10	10-20
T60 (in ms)	50-200	200-600	600-1000
Talker-to-ASV Distance (in cm)	10-50	50-100	100-150

Table 3.8: Parameter Settings for Replay Configurations in ASVSpooF 2019 Challenge Dataset. After [5].

	A	B	C
D_a (in cm)	10-50	50-100	>100
(Q)	perfect	high	low

3.2.3.2 Physical Access (PA)

The replay spoof speech signals are generated with various acoustic and replay configuration parameters. The grading of the acoustic configuration parameters, such as room size, reverberation time in T60, and talker-to-ASV distance, is shown in Table 3.7. Whereas, grading of the replay configuration parameters, such as attacker-to-talker recording distance (D_a) and loudspeaker quality (Q), is shown in Table 3.8. The statistics of the partition of the utterances into training, Dev, and Eval set, are shown in Table 3.5. The partition of the speakers is similar to that of LA scenario.

3.2.4 ASVSpooF 2021 Challenge Dataset

This dataset was released during a satellite event of INTERSPEECH 2021 for developing countermeasures for ASV systems [170]. This dataset aims to develop generalized countermeasures against LA, PA, and *DeepFake* attacks. Furthermore, the ASVSpooF 2021 dataset is partitioned into training, Dev, and Eval sets, where the training and Dev sets are the same as that of ASVSpooF 2019 dataset. However, the Eval set of ASVSpooF 2021 challenge contains new utterances. In particular, for LA scenario, Eval set utterances are generated by transmitting genuine utterances across VoIP networks. This results in the presence of coding and transmission artifacts, but no additive noise. The PA Eval set predominantly contains real replayed speech. However, a small proportion of simulated replay is also present. These factors are similar to that of ASVSpooF 2019, but are more comprehensive.

Furthermore, the ASVSpooF 2021 dataset additionally introduced speech Deep-

Fake data for the first time in the anti-spoofing literature. The DF evaluation data is a collection of bonafide and spoofed speech utterances processed with different lossy codecs, namely, mp3, m4a, and ogg, which are used typically for media storage. Audio data is encoded and then decoded to recover uncompressed audio.

The DeepFake dataset is generated using TTS and VC algorithms. The LA subset also uses the TTS and VC utterances, however the DeepFake utterances include compressed data (rather than telephony). The compression methods include *mp3* and *m4a*.

3.2.5 Biometrics: Theory, Applications, and Systems 2016 (BTAS 2016) Dataset

This dataset was released for the speaker anti-spoofing competition during IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2016) [7]. It considers all the major types of spoofing attacks, namely, replay, SS, and VC. ⁴BTAS 2016 dataset uses AVSpooF dataset [6]. Genuine utterances in BTAS 2016 dataset are recorded from the 44 subjects, which consists of 31 males and 13 females. The recording is performed in 4 sessions over the period of 2 months, with varying recording setups and environmental conditions. Three types of recording devices, namely, laptop using microphone AT2020USB+, Samsung Galaxy S4 phone, and iPhone 3GS are utilized for the genuine speech signal recording, which consists of 3 types: (1) reading part of 10 or 40 pre-defined sentences read by subjects (read), (2) pass-phrases part of 5 short prompts read by subjects (pass-phrases), and (3) free speech part of a free speech about any topic for 3 to 10 minutes (free). The details of the BTAS 2016 dataset recordings *w.r.t.* the session and recording type is shown in Table 3.9. The statistics of the dataset *w.r.t.* replay configuration is shown in Table 3.10. It can be observed that the training and Dev set consists of the similar kind of spoofing attack algorithms. Hence, known attacks are present in the Dev set. However, the test set consists of two unseen replay attacks, which are known as *unknown* attacks.

3.2.6 POp noise COrpus (POCO)

Since the past 08 years, significant amount of work has been done in the SSD literature for ASV task. However, detection of live speech has only been paid

⁴The BTAS 2016 dataset is available at <https://www.idiap.ch/en/dataset/avspooF> {Last Accessed: June 1, 2022}.

Table 3.9: Statistics of the BTAS 2016 Dataset *w.r.t.* the Session and Recording Type. After [6].

Recording type	Session 1	Session 2-4	Total
read	10 sentences	40 sentences	25.96 hours
pass-phrases	5	10	4.73 hours
free	≥ 5 min	≥ 3 min	38.51 hours

Table 3.10: Number of Utterances in BTAS 2016 Dataset. Acronyms in this Table Stands for the Following Terms: SS- Speech Synthesis, VC- Voice Conversion, RE- Replay, LP- Laptop, PH1- Samsung Galaxy S4 Phone, PH2- iPhone 3GS, PH3- iPhone 6S, HQ- High Quality Speakers. After [7].

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

- : Not Applicable

attention to recently, by using the recent standard corpora, POCO [8] ⁵. For liveness detection of speech, pop noise is utilized as a characteristic of live speech. Pop noise is produced due to the breathing effects captured by the microphone. If microphone in ASV system is *assumed* to be placed close to the genuine/live speaker, then it is able to capture the pop noise effectively. Therefore, pop noise becomes a suitable acoustic feature for distinguishing a live speech from a spoof (especially replayed) speech signal. To that effect, the POCO dataset is developed to investigate the voice liveness detection for ASV.

The POCO dataset consists of speech recordings of 66 speakers (32 male and 34 female) aged from 18 to 61 years, with varying levels of English language fluency and accent. The dataset is recorded with 22050 Hz sampling frequency and a

⁵The POCO dataset can be found at <https://github.com/aurtg/poco> {Last Accessed: June 1, 2022}.

bit-depth of 16-bits. The dataset is organized into three parts, namely, Recording with microphone A (RC-A), Eavesdropping (RP-A), and Recording with microphone array (RC-B). These parts differ from each other in number of microphones, type of microphone(s) used, and presence/absence of *pop filter*. The details of these 3 parts of the dataset are given in Table 3.11. The subset RC-A represents live speaker recordings having pop noise. The subset RP-A consists of emulated scenario of spoofed speech by using pop filter to eliminate/diminish pop noise. While RC-A and RP-A consists of speech data captured by a single microphone, the subset RC-B consists of speech data captured by an array of 15 microphones. Like the RC-A subset, the RC-B subset also doesn't use pop filter and hence, corresponds to the live speech. Speech signals in RC-B set are recorded in 3 settings *w.r.t.* speaker-microphone distances, namely, 5 cm, 10 cm, and 20 cm. The effect of human breath on the microphone depends on the uttered phoneme type. Thus, the POCO dataset is collected such that it consists of speech recordings of 44 words corresponding to 44 phonemes in the English language, as shown in Table 3.12.

Table 3.11: The Three Subsets of POCO Dataset. After [8].

Subset	Microphone Name	Microphone Directionality	Number of Microphones	Distance of Speaker from the Microphone (in cm)	Pop Filter
RC-A	Audio-Technica AT4040	Cardoid	1	10	No pop filter used
RP-A	Audio-Technica AT4040	Cardoid	1	10	TASCAM TM-AG1
RC-B	Audio-Technica AT9903	Omnidirectional	15	5, 10, and 20	No pop filter used

For each of the recording setting (RC-A, RP-A, RC-B (5 cm), RC-B (10 cm), and RC-B (20 cm)), each word shown in Table 3.12 was repeated 3 times by every speaker. Furthermore, in the case of RC-B setting, where multiple microphones were used, all the microphones were tuned independently so that the maximum volume remained below the threshold of -6 dB. The standard partition is not provided by the organizers and hence, experiments using this dataset can be conducted by considering non-overlapping training and testing subsets.

Table 3.12: The Set of Words Utilized in POCO Dataset. After [8].

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

Table 3.13: Microphone Array Settings for ReMASC Dataset. After [9].

Device	D1	D2	D3	D4
Model	Google AIY	Respeaker 4 Linear	Respeaker V2	Amlogic 113X1
Sample Frequency	44100	44100	44100	16000
Number of Channel	2	4	6	7
Bit Depth	16	16	32	16
Microphone Array Structure	2-Mic Linear	4-Mic Linear	6-Mic Circular	6-Mic Circular & 1 Central Mic

Table 3.14: Statistics of the ReMASC Dataset *w.r.t.* Various Acoustic Environments. After [9].

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

3.2.7 Realistic Replay Attack Microphone Array Speech Corpus (ReMASC)

ReMASC corpus is specifically designed to develop the CMs for VAs [9]⁶. There are important differences between ASV and VAs, primarily, the distance between the speaker and the microphone is larger in VAs. Furthermore, VAs utilize a microphone array as opposed to the single microphone in ASV. In the ReMASC dataset, 132 voice commands are used. These voice commands consists of 273 unique words for phonetic diversity. The number of speakers in the dataset are 50, among which 22 are female speakers, and 28 are male speakers. Furthermore,

⁶This dataset is publicly available at <https://github.com/YuanGongND/ReMASC> {Last Accessed: June 1, 2022}.

Table 3.15: Statistics of the Subset of the ReMASC Dataset Partitioned into Three Subsets. After [9].

	Training	Dev	Eval
Genuine	2820	924	3308
Spoof	7392	1884	9203
Total	10212	2808	12511

36 speakers are native speakers of English language, 12 are Chinese native speakers, and 2 are Indian speakers. The speech data is collected for 4 systems, details of which are shown in Table 3.13. Furthermore, to study the effect of recording device in replay attack, one low quality (iPod Touch (Gen5)), and one high quality recorder (Tascam DR-05) is used. It is observed that even with Tascam DR-05, channel and background noise are unavoidable. To that effect, for additional replay source recordings, Google TTS is used, which is free from transmission channel and background noise. For playback, 4 devices are used: A) Sony SRSX5, B) Sony SRSX11, C) Audio Technica ATH-AD700X headphone, and D) iPod Touch. Moreover, an additional playback device is used in the vehicular environment as the built-in vehicular audio system. The ReMASC data is recorded in 4 types of environments, namely, outdoor environment, vehicle environment, indoor environment-1, and indoor environment-2. The statistics of the dataset along with corresponding environments is shown in Table 3.14.

For this dataset, standard partition, protocols, and performance evaluation metrics are not provided by the dataset organizers. However, in this thesis, we have utilized the ReMASC dataset, which consists of ~ 25500 of utterances that are partitioned into three subsets, namely, training, Dev, and Eval sets. The corresponding statistics are shown in Table 3.15. Notably, the partition is *disjoint* in terms of the speakers and the data distribution among the environments is non-uniform. Various datasets designed for various spoofing attacks are summarized in Table 3.16.

3.3 Existing Feature Sets

In this thesis, results for the proposed feature sets are compared with existing state-of-the-art feature sets, such as CQCC, MFCC, LFCC, SECC, and TECC. The proposed feature sets are described in subsequent chapters. However, the various state-of-the-art feature sets, which are repeatedly explored along with the proposed feature sets for performance comparison, are described next:

Table 3.16: Summary of Various Datasets Utilized in this Thesis.

Dataset	Spoofing Attacks	Remark
ASVSpooof 2015	SS and VC	Vocoders and USS
BTAS 2016	SS, VC and Replay	Common Consumer Grade Devices
ASVSpooof 2017	Real Replay	Real Replay using Common Consumer Grade Devices
ASVSpooof 2019	SS, VC (LA)	Adavanced NN-based SS and VC,
	Simulated Replay (PA)	Simulated (Controlled) Replay
ASVSpooof 2021	DeepFake	Processed with Different Lossy Codecs
	LA	Channel and Compression Variability
	PA	Real, Variable Spaces
ReMASC	Real Replay	For VAs
POCO	Simulated Replay	VLD
VSDC	Multi-Order Replay	For VAs

3.3.1 CQCC

This feature set is derived from CQT, which is well known for the analysis of speech signal. In STFT, central frequencies of the subband filters are linearly-spaced, whereas in CQT, they are geometrically-spaced. The CQT is perceptually-motivated based on Weber’s law, which states that the change in a stimulus that will be just noticeable is a constant ratio of the original stimulus. When this law is applied to the perception mechanism of the sound, then it gives higher frequency resolution at lower frequency regions, and lower frequency resolution at high frequency regions [171]. CQT maintains a constant Q-factor for all the subband filters. Uniform sampling is performed on CQT followed by DCT to derive CQCC feature set [172]. The details of the CQT are discussed in Chapter 5.2.3, where it is used for VLD. The parameters of the CQT can be tuned based on the application. To develop the CM system for ASV, CQCC features are extracted by setting the maximum frequency to Nyquist rate, and minimum frequency at 15 Hz. The number of bins per octave is set to 96. Features are extracted with 30 DCT static coefficients (with log-energy) appended by Δ and $\Delta\Delta$, resulting in a total 90-D feature vector.

3.3.2 Cepstral Feature Sets

This feature set is extracted from the log-magnitude spectrogram followed by DCT. The Nyquist frequency range in log-magnitude spectrogram is extracted by using 257 frequency bins. 40 coefficients are retained as the static coefficients with Δ and $\Delta\Delta$ coefficients appended to it in order to form the 120-D cepstral feature

set.

3.3.3 MFCC and LFCC

MFCC is proved to be one of the most successful feature set in a wide range of speech technology applications including speech and speaker recognition. It also mimics the auditory representation, such as CQCC. The windowed speech signal is processed through Fourier transform to produce STFT. The weighted sum is performed for each Mel scale subband filter. Then, DCT is applied and desired number of cepstral coefficients are extracted to get MFCCs. In this thesis, we have used 40 Mel scale subband filters for feature extraction. 13-D and 40-D static coefficients have been extracted for various experiments. For the SSD task, LFCC is found to be a successful feature set, which is used as a baseline feature set for ASVSpooof 2019 challenge [5,76]. In case of LFCC, Mel scale filterbank is replaced by linear-scale filterbank, where central frequencies of the subband filters are linearly-spaced. LFCCs are extracted with 40 linear-scale subband filters. All 40 cepstral coefficients are retained and appended with Δ and $\Delta\Delta$ coefficients to form 120-D LFCC feature set.

3.3.4 TECC and SECC

In this thesis, enhanced-TEO and cross-TEO-based feature sets, namely, ETECC and CTECC are proposed for SSD task. Enhanced-TEO and cross-TEO are the modifications of TEO [173]. Hence, for fair comparison, the performance of the proposed feature sets are compared against TEO- and squared energy operator-based, TECC and SECC feature sets. Except for the energy estimation approach, the remaining procedure for the extraction of TECC and SECC feature sets is similar to that of ETECC feature set as explained in Chapter 4, Section 4.4.2.

3.4 Classifiers Used

3.4.1 Gaussian Mixture Model (GMM)

GMM is a type of clustering, where each cluster has a shape of Gaussian distribution. The probability density function (*pdf*) for a univariate Gaussian distribution is given [174]:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad (3.1)$$

where μ and σ are the mean and variance of a Gaussian distribution, respectively. Here, we have soft decisions, i.e., a certain probability is assigned to a particular data point for its belonging to a specific cluster. This means each data point could belong to any distribution with a corresponding probability. To estimate this type of model, Expectation Maximization (EM) algorithm is employed. The EM algorithm is exploited in GMM to find out the Maximum Likelihood Estimation (MLE) parameters for the given data. In EM algorithm, first the Gaussian distribution is utilized to obtain random clusters for initialization of the algorithm. After that, the probabilities of each data point for belonging to a particular cluster is calculated. Using this probability information, the clusters are re-estimated by updating their means and variances.

GMM for genuine (natural) speech (λ_n) is trained using genuine utterances, whereas GMM for spoofed speech (λ_s) is trained using spoofed utterances from the training set. Now, these trained GMMs are used for testing purpose. The final scores for the extracted features (X) of the test utterance, are calculated in terms of log-likelihood ratio (LLR) as follows:

$$LLR = \log(p(X|\lambda_n)) - \log(p(X|\lambda_s)), \quad (3.2)$$

where $p(X|\lambda_n)$ and $p(X|\lambda_s)$ are the likelihood scores obtained using GMM for genuine and spoofed utterances, respectively. The obtained scores help to classify whether the unknown sample belongs to the natural or spoofed class.

3.4.2 Support Vector Machine (SVM)

In Chapter-4 and Chapter-5, SVM is utilized as a classifier for experiments. SVM is a non-probabilistic binary linear classifier, as it assigns any new data point directly to one of the classes. The SVM gives an optimal hyperplane, given labeled training data, which categorizes new examples [174]. Consider a linear model given by:

$$z(x) = w^T \psi(x) + b, \quad (3.3)$$

where w represents the weight vector, $\psi(x)$ represents the fixed feature space transformation, and b represents the bias. We are interested to obtain all the data points correctly classified, i.e., $t_n y(x_n) > 0$. Here, x_n represents the n^{th} input data point, $y(x_n)$ represents the output, and t_n represents the corresponding target value, which takes value -1 and 1. If the data point is correctly classified, then $t_n y(x_n)$ will always be positive. Furthermore, the distance (d) between the data

point x_n , and decision hyperplane is given by [174] :

$$d = \frac{t_n z(x_n)}{\|w\|} = \frac{t_n(w^T \psi(x_n) + b)}{\|w\|}. \quad (3.4)$$

Now, the margin is given by the perpendicular distance of the hyperplane to the closest data point, x_n . Here, the motive is to optimize the parameter w and b in order to maximize this distance. Hence, the solution for maximum margin is given by [174] :

$$\arg \max_{w,b} \left[\frac{1}{\|w\|} \min(t_n(w^T \psi(x_n) + b)) \right]. \quad (3.5)$$

This optimization problem is further evaluated by use of Lagrange theorem to obtain the optimum hyperplane for classification purpose [174].

If the given data is not linearly separable, then the kernel trick is used for transformation of data into a suitable form for the classification task [174]. This transformation in SVM is motivated by the Cover's theorem, which states that given a set of training data that is not linearly separable, one can transform it into a training set that is linearly separable by projecting it into a higher-dimensional space via some non-linear transformation [175]. We have used 2-class linear kernel for the classification task. Furthermore, the regularization parameter instructs the SVM optimization about the maximum limit of misclassifying each training example, i.e., how much it can misclassify. For large values of regularization parameters, a smaller margin hyperplane will be chosen by the optimizer if that hyperplane does a better job of getting all the training points classified correctly. On the other hand, a very small value of regularization parameter will direct the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassified more points. In this thesis, L2 regularization is used along with hinge loss for "maximum-margin" classification [176].

3.4.3 Convolutional Neural Network (CNN)

CNN are a special type of neural networks that processes data in a grid topology [177]. Rather than the conventional matrix multiplication, CNNs use convolution operation in their structure. In deep learning, a convolution operation is applied on a multi-dimensional input array using a multi-dimensional *kernel*. The *Kernel* is kept smaller than the size of the input and slides over the entire input during its operation. In this way, convolution operation accomplishes parameter sharing resulting in the need of lesser parameters and hence, lesser memory requirement for a specific task. Let the features extracted from the speech signal is denoted

as $X \in \mathbf{R}^{f \times t \times c}$, where t , f , and c are time index, frequency index, and number of input channels, respectively, convolution is done using a weight matrix $W \in \mathbf{R}^{m \times m}$, which transforms the matrix into $X^1 \in \mathbf{R}^{(f-m+1) \times (t-m+1) \times c^1}$, where c^1 is the number of output channels. Generally, the convolution operation is followed by a pooling operation that reduces the variability arising in the input. Studies suggest that deep learning architectures are more capable to exploit the discriminative features and use them to get trained for classifying the unknown speech signal accurately. Hence, CNN is used as a classifier for VLD, SS, and VC-based SSD task.

3.4.4 Light-CNN

The LCNN architecture is employed in this thesis since it is one of the successful architectures for replay SSD task [17, 112]. LCNN architecture uses Max-Feature-Map (MFM) activation operation, which is a special case of max-out, for learning with a few parameters [178]. MFM utilizes competitive selection strategy, which plays the role of efficient feature selection. MFM function is defined as [178]:

$$y_{ij}^k = \max(x_{ij}^k, x_{ij}^{k+\frac{N}{2}}), \quad (3.6)$$

where the number of channels of the input convolution layer is $2N$, ($1 \leq k \leq N$), ($1 \leq j \leq W$), and ($1 \leq i \leq H$). Here, i and j indicate the feature component and frame number, respectively. Each convolution layer is a combination of two independent terms previously calculated from the input layer's output. The MFM activation function is used then to calculate element-wise maximum of those parts. Max-pooling layers with an optimum size of the kernel and stride is used for dimensionality reduction.

3.4.5 Residual Neural Networks (ResNet)

ResNets are one of the popular DNN-based classifiers and introduced to take the advantage of DNN by integrating the high/mid/low-level features. DNN architectures are generally facing the problem of vanishing/exploding gradients and could not learn the fine high-level features and hence, affecting the performance of the system. To alleviate this issue, ResNets are introduced, which utilizes the identity mapping as shown in Figure 3.1 [18]. In Figure 3.1, $F(y)$ represents the mapping of input signal y by the block of convolutional layers. Identity mapping allows to stack more number of layers without introducing the vanish-

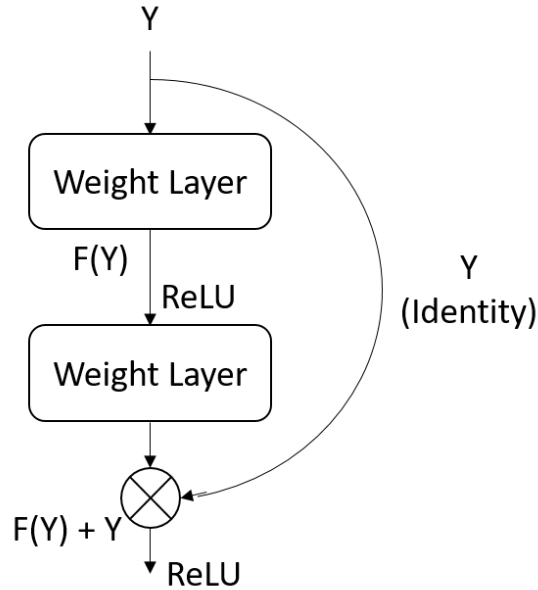


Figure 3.1: Residual Learning: A Building Block. After [18].

ing/exploding gradients and permits the possibility of smooth convergence. The increase in layers of DNN allow to learn high-level features and thus, improving the performance of the system. ResNets are utilized in this study as they are one of the successful architectures for SSD task in ASVSpooof 2019 and ASVSpooof 2021 challenge campaigns [108, 134, 142–145]. In this thesis, ResNet architecture is employed for pop noise detection and severity-level classification of the dysarthric speech.

3.5 Performance Evaluation Metrics

For evaluation of the SSD and VLD systems, various performance measures have been employed, namely, Equal Error Rate (EER), t-DCF, % classification accuracy, and Area Under the Curve (AUC). These performance metrics are briefly explained as follows:

3.5.1 Equal Error Rate (EER)

The LLR scores obtained from CM system is used to compute % EER. The EER is derived from the detection error trade-off (DET) curve, which represents the performance on detection tasks that involve the trade-off of error types [179]. In SSD task, there are two types of errors, i.e., false alarm rate ($P_{fa}(s)$) and miss rate

($P_{miss}(s)$). For arbitrary threshold s , these error rates are defined as:

$$P_{fa}(s) = \frac{\text{Number of spoofed trials with score } > s}{\text{Total number of spoofed trials}}, \quad (3.7)$$

$$P_{miss}(s) = \frac{\text{Number of genuine trials with score } \leq s}{\text{Total number of genuine trials}}. \quad (3.8)$$

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

$$EER = P_{fa}(s_{EER}) = P_{miss}(s_{EER}). \quad (3.9)$$

3.5.2 tandem - Detection Cost Function (t-DCF)

We also utilized the other performance metric called t-DCF, which allows the joint (tandem) evaluation of the SSD and ASV systems [30]. Hence, it will require the LLR score-values from SSD as well as ASV systems. As the ASV system scores for ASVSpooof 2019 challenge dataset are provided by the challenge organizers, we used t-DCF for the evaluation of this dataset. The t-DCF is the extension of the Detection Cost Function (DCF), which was used in the NIST challenges [180]. DCF requires the values of target priors, costs for missing the target (C_{miss}), false alarm (C_{fa}), and error probabilities (P_{miss} and P_{fa}) of ASV system. t-DCF takes into account all these parameters for both SSD and ASV systems. With cascade or parallel arrangement of SSD and ASV systems, there are six possible action pairs based on acceptance or rejection by the combined SSD and ASV systems. Proposition set represents actual states of the nature consists of three classes, i.e., target, non-target, and spoof speech utterances, which will have their own prior as π_{tar} , π_{non} , and π_{spooof} , respectively. Let C_{fa}^{asv} and C_{miss}^{asv} represents the cost of ASV system accepting a non-target trial and rejecting a target trial. Also, C_{fa}^{cm} and C_{miss}^{cm} represent the cost of SSD system accepting a spoof trial and rejecting a human trial, respectively. Error probabilities of the ASV and SSD systems are independent. Hence, their joint error probability is the multiplication of the two independent error probabilities. Let s and t be the thresholds for the SSD and ASV systems, respectively. Then, error probabilities of various errors can be obtained as [30]:

- SSD does not miss the genuine speech and ASV rejects the target:

$$P_a(s, t) = (1 - P_{miss}^{cm}(s)) \cdot P_{miss}^{asv}(t). \quad (3.10)$$

- SSD does not miss the genuine speech and ASV accepts the non-target:

$$P_b(s, t) = (1 - P_{miss}^{cm}(s)) \cdot P_{fa}^{asv}(t). \quad (3.11)$$

- SSD passes the spoof speech and ASV does not miss the target:

$$P_c(s, t) = P_{fa}^{cm}(s) \cdot (1 - P_{miss,spoof}^{asv}(t)). \quad (3.12)$$

- SSD misses the genuine speech:

$$P_d(s) = P_{miss}^{cm}(s). \quad (3.13)$$

By computing the priors, error probabilities and costs, t-DCF is defined as:

$$t - DCF(s, t) = C_{miss}^{asv} \cdot \pi_{tar} \cdot P_a(s, t) + C_{fa}^{asv} \cdot \pi_{non} \cdot P_b(s, t) + C_{fa}^{cm} \cdot \pi_{spoof} \cdot P_c(s, t) + C_{miss}^{cm} \cdot \pi_{tar} \cdot P_d(s). \quad (3.14)$$

The summary of t-DCF computation for ASV and CM system is systematically demonstrated in Figure 3.2.

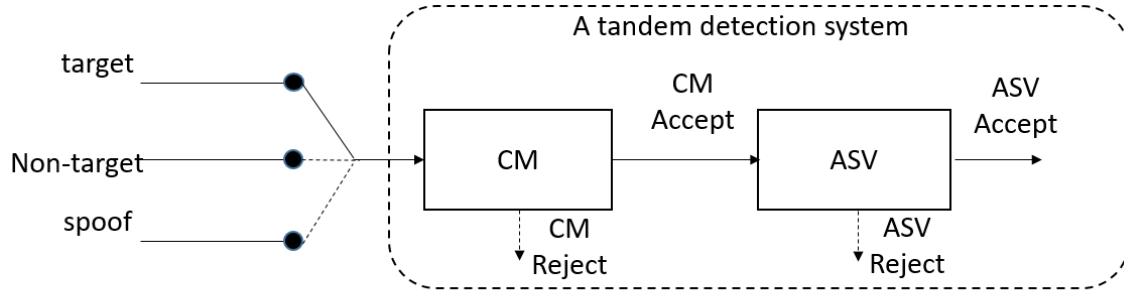
3.5.3 % Classification Accuracy

Percentage classification accuracy is a performance metric used to measure the number of data points classified correctly among all the test samples. In general terms, it is the estimation of the degree of closeness of a predicted value to that of the true value. Mathematically, % classification accuracy is defined as :

$$\% \text{ Classification Accuracy} = \frac{\text{Number of data points correctly classified}}{\text{Total number of data points}} \times 100. \quad (3.15)$$

3.5.4 Area Under the Curve (AUC) for Overlapping Region

Furthermore, the performance of our SSD systems can be evaluated using the area under the curve (AUC) for the overlapping region between the probability density functions (*pdfs*) of the LLR scores for genuine and spoof speech utterances, as shown in Figure 3.3. In this case, AUC provides an aggregate measure of performance of a model across all the possible classification thresholds.



Actual class	Tandem Decision	Unit cost	Actual class	Asserted Prior
Target	Reject	C_{miss}	Target	π_{tar}
Non-target	Accept	C_{fa}	Nontarget	π_{non}
Spoof	Accept	$C_{fa,spoof}$	Spoof	π_{spoof}
Target	Reject	C_{miss}		$\Sigma = 1$

$$t - DCF = \underbrace{C_{miss} \cdot \pi_{tar} \cdot P_a}_{\text{(CM accept, ASV reject)}} + \underbrace{C_{fa} \cdot \pi_{non} \cdot P_b}_{\text{(CM accept, ASV accept)}} + \underbrace{C_{fa,spoof} \cdot \pi_{spoof} \cdot P_c}_{\text{(CM accept, ASV accept)}} + \underbrace{C_{miss} \cdot \pi_{tar} \cdot P_d}_{\text{(CM reject)}}$$

Tandem Action → (CM accept, ASV reject) (CM accept, ASV accept) (CM accept, ASV accept) (CM reject)

Actual Class → Target Non-target Spoof Target

Figure 3.2: A Tandem System Consisting of ASV and SSD Modules is Evaluated using Three types of Trials: Targets, Nontargets, and Spoofing Attacks. Adapted from [30].

3.6 Score-Level Data Fusion

The score-level data fusion is performed on LLR scores (as shown in eq. (3.2)) obtained from the multiple systems in order to capture the possible complementary information. Score-level fusion of two systems using *linear weighted sum* is given as:

$$LLR_{fused} = \beta \cdot LLR_{S1} + (1 - \beta) \cdot LLR_{S2}, \quad (3.16)$$

where LLR_{S1} and LLR_{S2} are the LLR scores derived from the system $S1$ and system $S2$, respectively. The fusion parameter $\beta \in [0, 1]$ determines the contribution of each of the system during score-level fusion. This fusion can be performed by using well known Bosaris' toolkit, which provides a logistic regression solution to train a combination of weights for fusing multiple subsystems into a single sub-system, outputting a well-calibrated LLR [181]. To obtain the appropriate combination of weights, parameters were optimized as in the following mapping

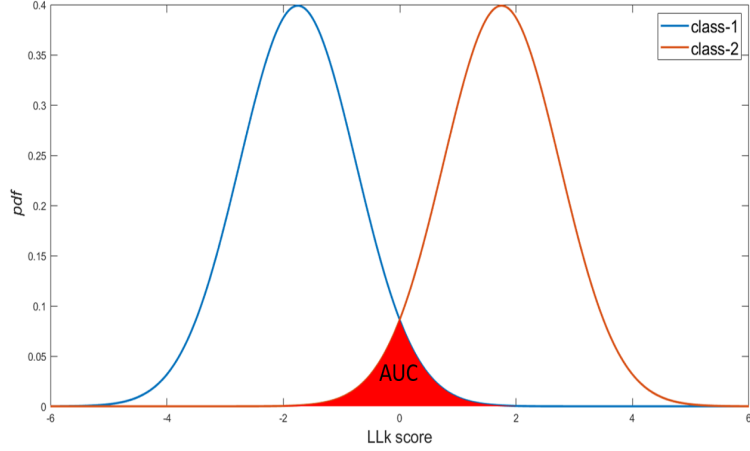


Figure 3.3: Demonstration of the AUC of the Overlapping Regions for the *pdfs* of the LLR Scores for the Two-class Classification Task. After [1].

function:

$$l_t = a + \sum_{i=1}^N b_i s_{it} + q_t' W r_t, \quad (3.17)$$

where l_t is the fused and calibrated output LLR for trial t , N is the number of sub-systems to be fused, s_{it} is the score of sub-system i for trial t , and q_t and r_t are optional quality vectors derived from the two sides (enroll, verify) of trial t . q_t' represents the transpose of the vector, q_t . The parameters to be optimized are the scalar offset a , the scalar combination of weights b_i , and the asymmetric matrix W , which effectively combines the two quality vectors into a quality score for the trial. The parameters are optimized using *logistic regression*, which minimizes an objective function.

The score-level fusion can be performed for two different systems having different feature sets and common classifier. This fusion can capture the possible complementary information in the feature sets for the intended task. Similarly, classifier-level fusion of the scores can be performed for the systems having different classifiers and common feature set and may be able to produce the relatively better performance than the standalone SSD system. This strategy can capture the possible complementary information in different classifiers.

3.7 Chapter Summary

In this chapter, various components of experimental setup, such as datasets, feature sets, classifiers, evaluation metrics, and score-level fusion techniques that are utilized in this thesis are discussed. The details of the data collection strategy, along with the statistics of the partition is provided. Furthermore, the brief tech-

nical details of the feature sets and classifiers utilized in this thesis are discussed. Finally, various evaluation metrics and score-level data fusion techniques are discussed. In subsequent chapters, several proposed feature sets are discussed. To that effect, experimental setup components required to validate the performance of the proposed feature sets can be referred from Chapter 3.

CHAPTER 4

Features using TEO

4.1 Introduction

This¹ chapter discusses the proposed handcrafted features, namely, ETECC, CTECC, and CFCCIF-ESA, which are derived using the concept of TEO and effectively utilized in this thesis for building the CMs against the major spoofing attacks. The ETECC feature set utilizes Enhanced Teager Energy Operator (ETEO), which is able to capture the high frequency energies more accurately by compensating the *signal mass*. To that effect, it effectively captures the energies in the mid and high frequency regions, where replay characteristics are present. Furthermore, CTECC_{max} feature set in this thesis captures the maximum relation distortion among the multi-channel subband filtered signals, and able to produce better performance than the other existing feature sets. TEO-based ESA algorithm is uti-

¹This Chapter is based on the following publications:

- **Ankur T. Patil**, Rajul Acharya, Hemant A. Patil, and Rodrigo Capobianco Guido, "Improving the potential of Enhanced Teager Energy Cepstral Coefficients (ETECC) for replay attack detection," in *Computer, Speech & Language*, Elsevier, vol. 72 (2022), pp. 101281.
- Rajul Acharya, Harsh Kotta, **Ankur T. Patil**, and Hemant A. Patil, "Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, June 2021, pp. 6364-6368.
- **Ankur T. Patil**, Rajul Acharya, Pulikonda Krishna Aditya Sai, and Hemant A. Patil, "Energy Separation-Based Instantaneous Frequency Estimation for Cochlear Cepstral Feature for Replay Spoof Detection," in *INTERSPEECH*, Graz, Austria, September 2019, pp. 2898-2902.
- **Ankur T. Patil**, Hemant A. Patil, and Kuldeep Khorja, "Effectiveness of Energy Separation-Based Instantaneous Frequency Estimation for Cochlear Cepstral Features for Synthetic and Voice Converted Spoofed Speech Detection," in *Computer, Speech & Language*, Elsevier, vol. 72 (2022), pp. 101301.
- **Ankur T. Patil**, Anand Therattil, and Hemant A. Patil, "On Significance of Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants," submitted in *Computer, Speech & Language*, Elsevier, July 2022.

lized in CFCCIF framework to estimate the IFs. The subsequent section of this chapter includes motivation of TEO, derivation of TEO, and ESA followed by the details of the proposed feature sets and their performance for SSD task.

4.2 Motivation for TEO

In the traditional signal processing literature, signal energy is estimated by using the square operation (i.e., L^2 norm) over the entire signal under analysis, producing a *scalar*. This approach obviously fails to capture the existing speech nonlinearities including the properties of airflow pattern in the vocal tract system. To overcome this issue, TEO was proposed in [182]. TEO is a nonlinear differential operator, which accounts for the energy of the system required to generate a signal. Furthermore, the concept of TEO is extended to separate the amplitude modulation and frequency modulation in speech signals with high time resolution [31, 183]. The simple structure of the discrete-time version of TEO greatly reduces the time complexity in computation. These favorable properties of TEO motivated to develop TECC, which have been adopted for speech recognition task for normal *vs.* whispered speech [40, 184]. Furthermore, TECC has been considered one of the best performing feature set for the recognition of whispered speech [184]. In addition, TECC has also been useful to capture speech reverberation [106], which motivated its application for replay SSD task. The major contribution of this thesis is development of the three new feature sets, namely, ETECC, CTECC_{max}, and CFCCIF-ESA, which are based on the concept of TEO and these features have been successfully employed for the SSD task.

Recently, a mechanical mass analogy for digital signals was introduced in the literature, modifying TEO to produce the exact representation of signal's energy, as shown in the original study reported in [185]. The results described in that scientific piece of work clearly show that the introduced ETEO provides a better estimate of signal energy in comparison with the original TEO. Subsequently, the associated concepts of ETEO along with subband filtered speech signals have been successfully exploited to produce a specialized set of features known as ETECC for replay SSD task [10, 32]. The replay mechanism characteristics is supposed to exhibit in mid- and high-frequency regions and hence, proposed ETECC feature set is one of the successful candidate for replay SSD task.

Furthermore, a modified version of TEO for multi-channel signals, i.e., cross-TEO (CTEO) was utilized for replay SSD for VAs [186]. Modern VAs make use of microphone arrays, which help in better sound source identification using direc-

tivity cues (i.e., exploiting the *spatial* diversity). In such cases, CTEO was developed in [186], to select the appropriate subband channel for low noise compensation to improve the performance of the Automatic Speech Recognition (ASR) systems [187]. The speech signal recorded by a playback device contains distortions due to the intermediate devices. In [188], it is shown that a replayed speech signal shows lower damping as compared to its genuine speech because of the distortions introduced due to replay configurations. As a result of such distortions in replay signals, spectral spread is observed in the frequency-domain. Hence, the key idea of using a multi-channel energy tracking scheme using TEO was investigated, where the most noisy channels are selected via CTEO-based feature set, referred to as $CTECC_{max}$. The key idea here is to select the most noisy subband channel (as opposed to the least subband channel for speech recognition task [187]) to track maximum distortions in the transmission channel due to replay conditions. Thus, the greater the distortions in the speech signal, the more likely it is to be a replayed signal. It is the first study of its kind to exploit maximum distortion due to replay noise as a discriminative feature using CTEO for SSD task on VAs [11].

The CFCCIF-ESA feature set is proposed to combine the envelope (magnitude) information along with the phase information (in the form of IF) to effectively detect the SS, VC, and replay SSD. Previously proposed CFCCIF feature set composed of the information obtained from the magnitude envelope derived using cochlear filterbank and instantaneous frequency (IF), which is derived from Hilbert transform-based approach. However, this approach requires a speech segment of 10-30 ms and thus, it limits time resolution of IF estimation and hence, defeats the key objective of IF estimation to be able to fit the frequency of a sinusoid (corresponding to a monocomponent signal) locally and almost instantaneously [189]. Whereas, TEO-based ESA is known to accurately estimate the amplitude- and frequency-modulation patterns due to their relatively low computational complexity, high time resolution, and instantaneously adapting nature. To that effect, the ESA is exploited in CFCCIF framework instead of Hilbert transform to estimate the IFs of the subband filtered signal using cochlear filterbank, and consequently, it forms the structure of CFCCIF-ESA feature set.

The chronological development of the proposed feature sets are illustrated in brief (as discussed for each feature set in this Section) in Figure 4.1. The next Section of this chapter explains the mathematical description of the TEO. Furthermore, subsequent Sections explain the functionality of the proposed ETECC, $CTECC_{max}$, and CFCCIF-ESA feature sets along with related experimental setup

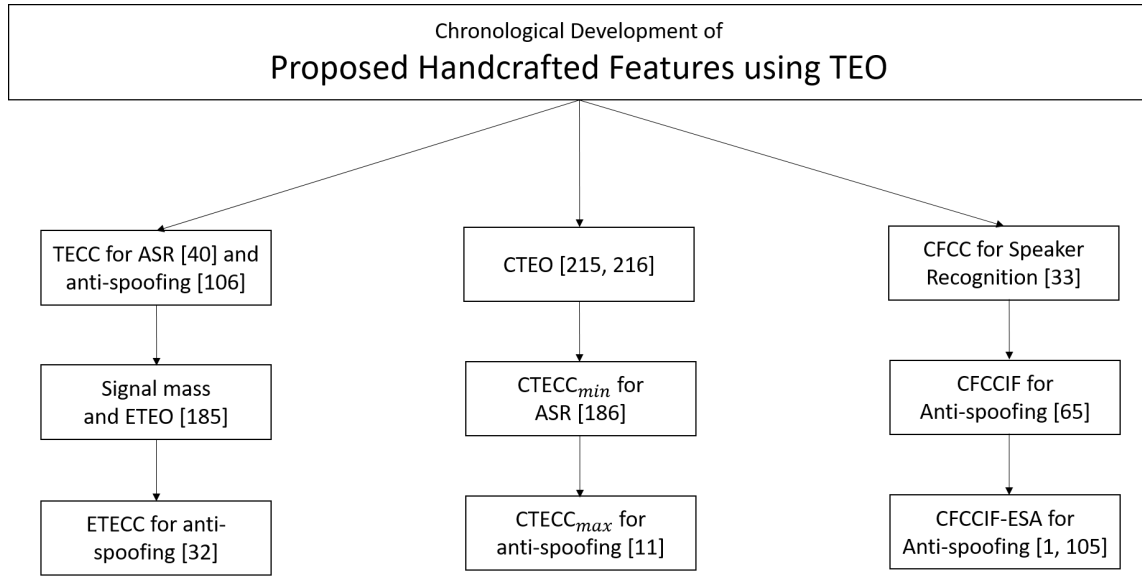


Figure 4.1: Brief Illustration of the Chronological Development of the Proposed TEO-based Features: ETECC, CTECC, and CFCCIF-ESA.

and results.

4.3 Derivation of TEO and ESA

Energy serves as link between speech production and perception because hearing is the process of detecting energy [190,191]. It was studied that the speech fine structures could not be extracted using the Fourier analysis [192]. This issue is alleviated using the TEO, which estimate the energy of the harmonic oscillator. The oscillating mass m suspended by spring with stiffness k forms the Simple Harmonic Motion (SHM). The dynamics describing SHM is given by 2^{nd} order ordinary differential equation, $\frac{d^2y}{dt^2} + \frac{k}{m}y = 0$, and whose solution is given by:

$$y(t) = A \cos(\omega t + \theta), \quad (4.1)$$

where A is amplitude, ω is frequency, and θ is phase. It can be observed that solution for SHM (i.e., $y(t)$) is a monocomponent signal. The energy (E) of a spring-mass system is directly proportional to the squared multiplication of its amplitude by its frequency of oscillations. Particularly,

$$E = \frac{1}{2}mA^2\omega^2, \quad \text{i.e., } E \propto A^2\omega^2, \quad (4.2)$$

where m , A , and ω correspond to the mass of the suspension, to the amplitude, and to the frequency of oscillations, respectively. The continuous-time version of TEO for signal $y(t)$ is given by [182]:

$$\psi\{y(t)\} = \left(\frac{dy(t)}{dt}\right)^2 - y(t) \cdot \frac{d^2y(t)}{dt^2}, \quad (4.3)$$

$$\psi\{y(t)\} = \dot{y}^2(t) - y(t) \cdot \ddot{y}(t). \quad (4.4)$$

where $\dot{y}(t)$ and $\ddot{y}(t)$ represents the single and double derivative of the signal $y(t)$.

In [182], AM-FM signals are modeled analogously to the mass-spring systems with the help of TEO, which is adopted to get the running estimate of signal's energy [31]. The non-linear modelling of the human speech production as in [192], [193], and [194], has been adopted for modelling and detecting modulations in speech resonances [42]. To derive the expression of the TEO, let us consider the discrete-time signal $y(n) = A \cos(\omega n + \theta)$, which represents the SHM. Furthermore, $y(n-1) = A \cos(\omega(n-1) + \theta)$ and $y(n+1) = A \cos(\omega(n+1) + \theta)$ represents the immediate past and future samples of the signal $y(n)$, respectively. Using the trigonometric identity for the arguments c and d ,

$$\cos(c+d) \cdot \cos(c-d) = \frac{1}{2}[\cos(2c) + \cos(2d)]. \quad (4.5)$$

We obtain,

$$y(n+1) \cdot y(n-1) = \frac{A^2}{2}[\cos(2\omega n + 2\theta) + \cos(2\omega)]. \quad (4.6)$$

Furthermore, we have:

$$\cos(2c) = 2 \cos^2(c) - 1 = 1 - 2 \sin^2(c). \quad (4.7)$$

From eq. (4.6) and trigonometric identity in eq. (4.7), we get:

$$y(n+1) \cdot y(n-1) = A^2 \cos^2(\omega n + \theta) - A^2 \sin^2(\omega) = y^2(n) - A^2 \sin^2(\omega). \quad (4.8)$$

Hence,

$$A^2 \sin^2(\omega) = y^2(n) - y(n+1) \cdot y(n-1). \quad (4.9)$$

If we restrict the value of $\omega < \pi/2$ in eq. (4.9), then $\sin(\omega) \approx \omega$ (i.e., this approx-

imation holds good for lower frequency region of the spectrum) and hence:

$$A^2\omega^2 \approx y^2(n) - y(n+1) \cdot y(n-1). \quad (4.10)$$

As suggested in eq. (4.2), the LHS of eq. (4.10) is nothing but energy E of the system to generate the signal. Thus, RHS of eq. (4.10) can be utilized to estimate the instantaneous energy of the signal $y(n)$, which is known as TEO and represented as $\psi\{y(n)\}$ [182]. Hence, using eq. (4.2), eq. (4.10), and representation of TEO, we have [182]:

$$\psi\{y(n)\} = y^2(n) - y(n-1) \cdot y(n+1) = A^2 \sin^2(\omega) \approx A^2\omega^2. \quad (4.11)$$

From eq. (4.11), it can be observed that TEO can produce positive as well as negative values. To alleviate this issue, we took absolute value of the TEO profile so that we obtain the positive energy values.

From eq. (4.11) and eq. (4.2), ideally TEO is developed only for monocomponent signals (moreover, solution of SHM is also a monocomponent signal). Furthermore, TEO possess time invariant property, which can be proved as follows.

For the system given in eq. (4.11), let us consider any arbitrary input $y_1(n)$ and any shift by k samples.

$$v_1(n) = \psi\{y_1(n)\} = y_1^2(n) - y_1(n-1) \cdot y_1(n+1). \quad (4.12)$$

Let us consider the second input obtained by shifting $y_1(n)$ by k samples, i.e.,

$$y_2(n) = y_1(n-k). \quad (4.13)$$

The output corresponding to the input $y_2(n)$ is:

$$v_2(n) = \psi\{y_2(n)\} = y_1^2(n-k) - y_1(n-k-1) \cdot y_1(n-k+1). \quad (4.14)$$

Similarly, from eq. (4.12):

$$v_1(n-k) = y_1^2(n-k) - y_1(n-k-1) \cdot y_1(n-k+1). \quad (4.15)$$

Comparing eq. (4.14) and eq. (4.15), we see that $v_2(n) = v_1(n-k)$. Hence, TEO is time invariant. Hence, the feature sets derived using TEO are shift (time) invariant.

To estimate the individual contribution of amplitude A and frequency ω to

the total energy, $\psi\{y(n)\}$, Energy Separation Algorithm (ESA) was developed, as in [31]. To understand the development of ESA in brief, let us first consider the real-valued continuous-time AM-FM signal is given by [42]:

$$\begin{aligned} y(t) &= a(t) \cdot \cos(\phi(t)), \\ &= a(t) \cdot \cos(\omega_c t + \omega_m \int_0^t p(\lambda) d\lambda + \theta), \end{aligned} \quad (4.16)$$

where $a(t)$ represents a time-varying amplitude signal modulated by the high frequency signal $\cos(\cdot)$, which results in AM. The time-varying instantaneous frequency (ω_i) is given by [31, 42]:

$$\omega_i(t) = \frac{d}{dt}\phi(t) = \omega_c + \omega_m \cdot p(t), \quad (4.17)$$

where $|p(t)| \leq 1$, ω_m corresponds to a maximum frequency deviation from ω_c , and θ is a phase offset. Applying TEO to the signal $y(t)$ results in,

$$\psi\{y(t)\} \approx a^2(t) \cdot \omega_i^2(t). \quad (4.18)$$

As discussed earlier, TEO is developed only for *monocomponent* signals. However, speech signals can be considered as the mixture of multi-component resonances, due to various cavities in the vocal tract system. Hence, bandpass filtering can be applied on speech signal to approximate the subband filtered signal as a monocomponent signal. Bounds are derived for the approximation errors for these subband filtered signals, which are negligible under general realistic conditions [194].

The instantaneous amplitude, $a(t)$, and instantaneous frequency, $\omega_i(t)$, are estimated using ESA [31]. Let,

$$\dot{y}(t) = \dot{a}(t) \cos(\phi(t)) - a(t)\omega_i(t) \sin(\phi(t)). \quad (4.19)$$

To make eq. (4.18) a valid approximation, let us assume two constraints:

- $a(t)$ and $p(t)$ are bandlimited with the highest frequencies ω_a and ω_p , respectively, and $\omega_a, \omega_p \ll \omega_c$.
- $\omega_a^2 + \omega_m \omega_p \ll (\omega_c + \omega_m)^2$

With the above two constraints, we have,

$$\psi\{\dot{y}(t)\} \approx \psi\{a(t)\omega_i(t) \sin(\phi(t))\} \approx a^2(t) \cdot \omega_i^4(t). \quad (4.20)$$

By combining eq. (4.18) and eq. (4.20), we get [42]:

$$|a(t)| \approx \frac{\psi\{y(t)\}}{\sqrt{\psi\{\dot{y}(t)\}}}, \quad (4.21)$$

$$\omega_i(t) \approx \sqrt{\frac{\psi\{\dot{y}(t)\}}{\psi\{y(t)\}}}, \quad (4.22)$$

where $\dot{y}(t)$ represents the first-order derivative of the signal, $y(t)$. The ESA is applicable under the constraints that $y(t)$ is a narrowband signal. For discrete-time systems, many variants of ESA are derived considering various approximations of derivative operator and AM-FM signal. One of the popular Discrete-time Energy Separation Algorithm (DESA) employs backward difference method, i.e., $\frac{dy(t)}{dt} \approx y(n) - y(n-1)$, (a.k.a. DESA-1a algorithm, where '1' and 'a' corresponds to single sample difference and asymmetric difference, respectively) and is given by [31]:

$$|a(n)| \approx \sqrt{\frac{2\psi\{y(n)\}}{1 - \left(1 - \frac{\psi\{y(n) - y(n-1)\}}{2 \cdot \psi\{y(n)\}}\right)}}, \quad (4.23)$$

$$\omega_i(n) = \arccos \left[1 - \frac{\psi\{y(n) - y(n-1)\}}{2\psi\{y(n)\}} \right]. \quad (4.24)$$

These algorithms exploit TEO to estimate the instantaneous amplitude and frequency components of the signal under consideration. As speech signals are modeled as a cascade of multi-component resonances, due to various vocal tract cavities, bandpass filtering usually precede the application of TEO. Eq. (4.11) provides a good estimate of signal energy only when the approximation $\sin(\omega) \approx \omega$ holds true.

Based on the concept of TEO, the ETEO, and CTEO were developed and these TEO-derived representations are effectively utilized in this thesis for replay SSD task. Furthermore, IFs estimated using ESA are utilized in CFCCIF-ESA feature representation, which performs relatively better against SS-, VC-based, and replay spoofing attacks. Each of these proposed feature sets are explained in the subsequent Sections.

4.4 ETECC Feature Set

4.4.1 Signal Mass (ρ) and ETEO

In order to obtain an exact estimate of signal's energy, ETEO was originally developed in [185]. In particular, the existing TEO was modified, adopting the new concept of signal mass. It is argued that all the discrete-time signals have a specific mass in association with them. This mass represents the resistance offered by the signal source to oppose its inertia, and only stationary signals keep their mass constant over time. Thus, if inertia of the signal is low, then it might be able to oscillate with higher frequency and vice-versa. Hence, making the use of signal mass, the exact signal energy (i.e., $A^2\omega^2$) can be obtained. Notably, signal mass can be completely derived in terms of $y(n)$ as discussed next. TEO (ψ) applied to any discrete-time signal $y(n)$ gives:

$$\lambda = \psi\{y(n)\} = y^2(n) - y(n-1) \cdot y(n+1) = A^2 \sin^2(\omega), \quad (4.25)$$

$$\lambda = A^2\omega^2 \frac{\sin^2(\omega)}{\omega^2}, \quad (4.26)$$

$$\lambda = A^2\omega^2 \operatorname{sinc}^2(\omega) = \frac{1}{2}\rho A^2\omega^2. \quad (4.27)$$

The comparison between eq. (4.2) and eq. (4.27), shows that ρ is the mass analogy found for a signal [106]. It is variable and a function of ω . This mass is constant for stationary signals, while it is variable for non-stationary signals. For signals with a time-varying frequency, mass is inversely proportional to the frequency of oscillation. When the air interacts with the cavities of the vocal tract system, speech is produced due to resonance [42]. Such cavities accentuate certain frequencies while attenuate the others [31]. Because of such variation in frequency distribution of speech signals, a variable signal mass exists. From eq. (4.27), ETEO, represented by Ψ , is given by:

$$\Psi = \frac{\lambda}{\frac{1}{2}\rho} = A^2\omega^2, \quad (4.28)$$

where ρ is obtained based on the following procedure [185]:

$$y(n-1) + y(n+1) = A[\cos(\omega n + \theta - \omega) + \cos(\omega n + \theta + \omega)], \quad (4.29)$$

$$\therefore \cos \omega = \frac{y(n-1) + y(n+1)}{2y(n)}. \quad (4.30)$$

Thus,

$$\omega = \arccos \left[\frac{y(n-1) + y(n+1)}{2y(n)} \right]. \quad (4.31)$$

Using the eq. (4.31), $\text{sinc}^2(\omega)$ can be derived to result in [185]:

$$\frac{1}{2}\rho = \text{sinc}^2 \omega = \begin{cases} 1, & \text{for } y(n) = 0, \\ \text{sinc}^2 \left(\arccos \left(\frac{y(n-1)+y(n+1)}{2y(n)} \right) \right), & \text{for } y(n) \neq 0, \\ \left(\frac{k^2+k\sqrt{k^2-1}-1}{|k+\sqrt{k^2-1}| \cdot \ln(|k+\sqrt{k^2-1}|)} \right)^2, & \text{for } |k| > 1, \end{cases} \quad (4.32)$$

where $k = \frac{y(n-1)-y(n+1)}{y(n)}$. Using eq. (4.32), signal mass can be estimated for each of the subband signals. Both λ and Ψ present a linear order of complexity in relation to length of the signal under analysis [185].

Figure 4.2-(a) shows an AM-FM signal and its corresponding estimated signal mass, where a synthetic AM-FM signal is considered as in [31], being defined as $[1 + 0.5 \cdot \cos(n\pi/50)] \cdot \cos(n\pi/5 + 4\sin(n\pi/100 + \pi/4))$. Notably, signal mass is low in regions, where the signal frequency is high and vice-versa. This is because, the higher the frequency is, the lower $\text{sinc}^2\omega$ is and hence, the smaller the signal mass is. At the same time, Teager energy is relatively high for high frequency regions, whereas enhanced Teager energy is much higher, as shown in Figure 4.2-(b).

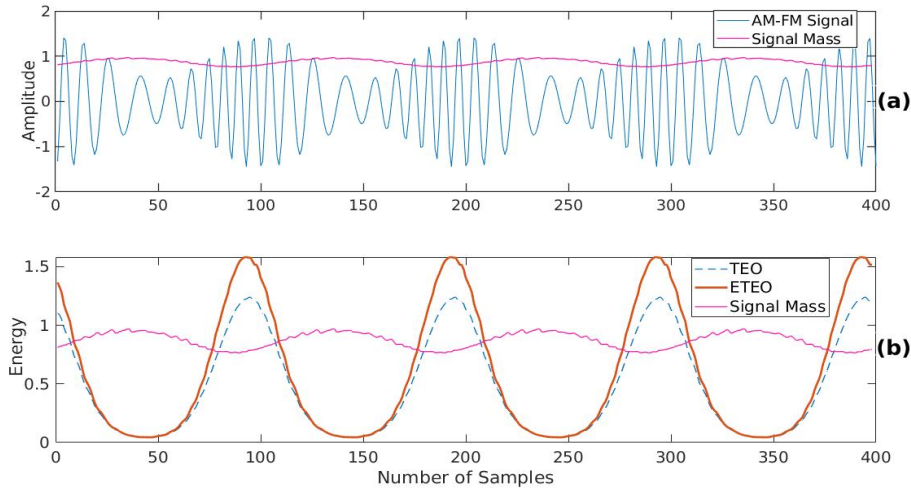


Figure 4.2: (a) A Synthetic AM-FM Signal Along With Superimposed Signal Mass, and (b) TEO, ETEO Profile Along With Signal Mass for the Signal Shown in Figure 4.2-(a). After [10,31].

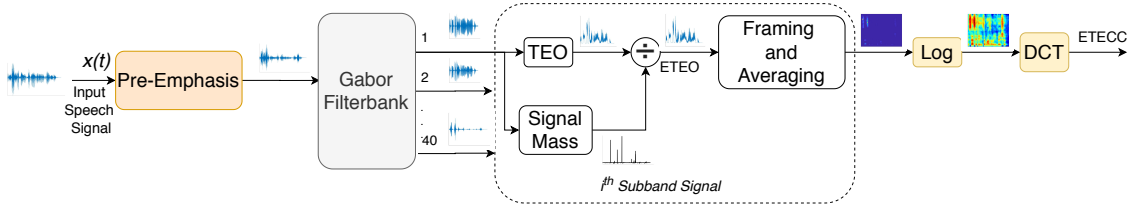


Figure 4.3: Functional Block Diagram of the Proposed ETECC Feature Set. After [32].

Algorithm 1 MATLAB Pseudo Code of Proposed ETECC Feature Set Extraction. After [10].

1. $x = \text{filter}([1 - 0.97], 1, x)$, pre-emphasis on speech signal x ,
2. $fbankG = \text{Gabor_fbank}(Q, bw)$, construct the Gabor filterbank with Q subband filters with bandwidth ' bw ',
3. **for** $i = 1 : Q$ **do**,
 - $y(i, :) = \text{filter}(fbankG)(i, :), 1, x)$, subband filtering using i^{th} subband filter,
 - $T_{subband}(i, :) = \text{TEO}(y(i, :))$, estimate energy using TEO,
 - $M_{subband}(i, :) = \text{signal_mass}(y(i, :))$, estimate the signal mass,
 - $E_{subband}(i, :) = T_{subband}(i, :)/M_{subband}(i, :)$, estimate energy using ETEO,
 - $E_{frames} = \text{enframe}(E_{subband}(i, :), win_len, win_shift)$, framing with appropriate window length and window shift,
 - $E_{avg}(i, :) = \text{mean}(E_{frames})$, Averaging over each frame,
- end for**
4. $E_{log} = \log(\text{abs}(E_{avg}))$,
5. $E_{static} = \text{DCT}(E_{log})$, static coefficients,
6. $E_{\Delta} = \text{delta}(E_{static})$, velocity coefficients,
7. $E_{\Delta\Delta} = \text{delta}(E_{\Delta})$, acceleration coefficients,
8. $ETECC = [E_{static}; E_{\Delta}; E_{\Delta\Delta}]$, ETECC feature set.

4.4.2 ETECC Feature Extraction

The functional block diagram for the ETECC feature extraction strategy, as in paper [32], is shown in Figure 4.3. Furthermore, MATLAB pseudocode for the same is illustrated in Algorithm 1. Spoofed replay signals can be mathematically expressed as the convolution of the genuine speech with the disturbances produced due to intermediate recording and playback channels, which are *band-pass* in nature [195]. To take advantage of this fact for replay SSD task, the speech signal is passed through a pre-emphasis filter for which the system function is $(1 - 0.97z^{-1})$, where z is the Z-transform variable. The pre-emphasized signal is, subsequently, passed through a Gabor filterbank having linearly-spaced subband filters [195–197].

The sound wave, i.e., the set of acoustic vibrations, is collected by pinna - a

part of the outer ear. Three tiny bones in the middle ear, namely, malleus, incus, and stapes, transform the acoustic vibrations into mechanical ones. The basilar membrane (BM) in the cochlea consists of fluid, which produce travelling wave in the BM due to mechanical vibrations. Particular frequency bands in sounds are sensed by a specified region of the BM. The travelling waves sweep from the base toward the apex of the cochlea and achieves a peak in the specified region, which depends on the sound frequency. Higher frequencies are sensed by the outer portion of the BM, and frequencies goes on decreasing as we move towards the inner core of the cochlea. This physiological structure motivates us to model the auditory system by using a subband filtering approach. The impulse response of the travelling wave can be modeled by the function $g(t) \in L^2(R)$ (i.e., Hilbert space of finite energy signals), which satisfies the following conditions:

- it should be the zero average function, i.e.:

$$\int_{-\infty}^{+\infty} g(t)dt = 0 \Rightarrow G(\omega)|_{\omega=0} = 0, \quad (4.33)$$

where $G(\omega)$ is Fourier transform of the $g(t)$;

- it suggests that $G(\omega)$ is bandpass in nature. The bandpass nature of the filter also helps to approximate the numerical computation [198];
- it also ensures the existence of a number C_g such that $C_g = \int_0^{+\infty} \frac{|G(\omega)|^2}{\omega} d\omega < +\infty$, which satisfies the *admissibility condition* in the original wavelet literature [199,200];
- it decreases to zero on both the ends. Similar nature is observed in psycho-acoustic experiments with the BM [201]. The impulse response of the second subband filter in the filterbank is shown in Figure 4.4.

Gabor filterbank satisfies all the above mentioned conditions to represent the impulse response of an auditory system [202]. Since the Fourier transform of a Gaussian function is a Gaussian, the impulse and frequency responses of Gabor filter are given by [31]:

$$g(t) = \exp(-a^2t^2) \cdot \cos(\omega_c t), \quad (4.34)$$

and

$$G(\omega) = \frac{\sqrt{\pi}}{2a} \left(\exp \left[-\frac{(\omega - \omega_c)^2}{4a^2} \right] + \exp \left[-\frac{(\omega + \omega_c)^2}{4a^2} \right] \right), \quad (4.35)$$

where ω_c represents the center frequency of a Gabor filter, and a controls its bandwidth. The frequency scale of the filterbank can be chosen either Equivalent Rectangular Bandwidth (ERB), Mel or linear depending upon the application.

The center frequencies and bandwidths for ERB or Mel scale filterbank emulate the filter structure in the human auditory system. For those filterbanks, center frequency and edges of the subband filters are linearly-spaced in lower frequency regions, whereas they are logarithmically increasing for the frequencies above 1 kHz. For linear scale filterbank, center frequencies and edges of the subband filters are linearly-spaced for the entire frequency range. In the proposed ETECC feature set, linear-scale Gabor filterbank is used by keeping the constant bandwidth for all the subband filters. This arrangement may help to estimate the reliable spectral information as it has the constant resolution across entire frequency range.

Gabor filter is a smooth function with a compact support and also possesses the optimal joint time-frequency resolution. The compact support of Gabor filter does not allow noise and distortions, present at distinct locations, to interfere with the Gabor filter in either time or frequency-domain. The optimal criteria for joint time-frequency resolution is achieved by minimum time-bandwidth product (TBP) dictated by Heisenberg's uncertainty principle, in signal processing framework [200]. In particular, the temporal variance (σ_t^2) and the frequency variance (σ_ω^2) of a signal, $x(t) \in L^2(\mathbb{R})$, satisfy

$$\sigma_t^2 \cdot \sigma_\omega^2 \geq 1/4. \quad (4.36)$$

Above inequality is obtained using Plancherel identity and Cauchy-Schwartz inequality. The theoretical proof of Heisenberg's uncertainty principle is provided in APPENDIX A. Eq. (4.36) becomes equality, when $x(t)$ is a Gaussian [200]. The product term, $\sigma_t^2 \cdot \sigma_\omega^2$ in eq. (4.36) is referred to as TBP. TBP indicates the area of Heisenberg's box, which represents the "richness" of signal information. Gabor filter has an important frequency response characteristic: the linear phase. Thus, it keeps the pattern or shape of the filtered output speech signal intact.

Subband filtering, as explained above, emulates the impulse response of the cochlea. Different regions of the BM respond to different frequency contents of the speech signal. The incident wave causes displacement of the inner hair cells to initiate neural activity responsible for the sound perception. The inner hair cells generate the neural activity in a single direction. This single direction movement can be represented by an energy function. In our earlier work, single direction movement is represented by the square of the filterbank output [105], however,

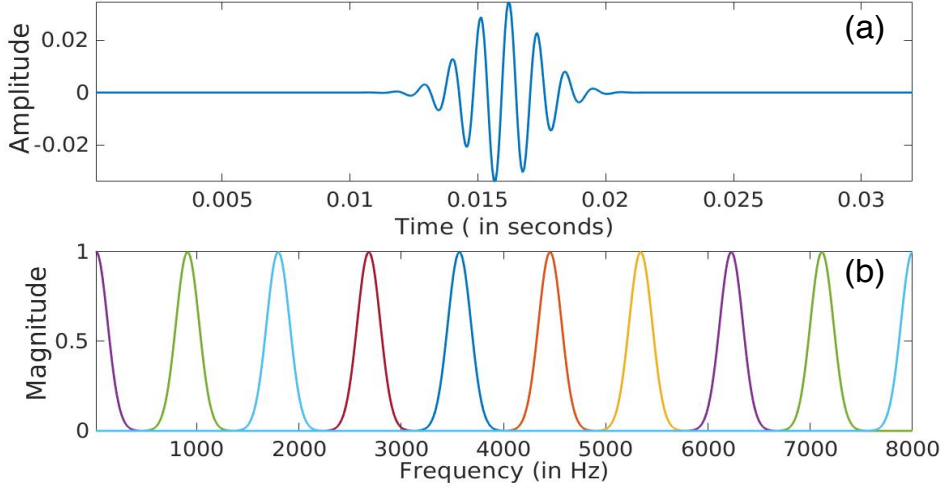


Figure 4.4: (a) Impulse Response of 2nd a Bandpass Filter in the Bank of 10 Sub-band Filters in a Gabor Filterbank with Linearly-Spaced Center Frequencies Between 0 Hz to 8 kHz, and (b) Frequency Response of a Gabor Filterbank.

it can be replaced by the other energy measures, such as TEO or ETEO, for more accurate estimation of the energy. In our proposed algorithm, ETEO is used as an energy measure at the filterbank output to mimic the single direction movement of the inner hair cells of the cochlea.

In the proposed framework, ETEO profile is estimated as given in eq. (4.28) for each subband filtered output. The numerator of eq. (4.28) represents the TEO profile, whereas the denominator consists of signal mass, which is, as mentioned, the key term for exact estimation of the energy. The unwanted glitches and spikes in the signal mass representation are eliminated by third-order one-dimensional median filter. The traveling wave in BM stimulates inner hair cells to generate the nerve impulses. This physiology can be modeled as estimating the energy of the subband filtered output, which can be referred to as hair cell output, expressed as [198]:

$$H(i, b) = \Psi(s(i, b)), \quad (4.37)$$

where $s(i, b)$ represents the b^{th} output sample of i^{th} a subband filter in a Gabor filterbank, and $\Psi(\cdot)$ represents the ETEO operator in eq. (4.28).

The inner hair cell output in cochlea is then transformed to an electrical signal, which is carried by the auditory nerves to the brain [203]. Its intensity can be modeled by Nerve Spike Density (NSD), which is computed by enframing and

averaging short frames of 25 *ms* with a frame shift of 10 *ms*, i.e.,

$$NSD(i, j) = \frac{1}{l} \sum_{b=n}^{n+l-1} H(i, b), \quad n = 1, N, 2N, \dots; \forall i, j, \quad (4.38)$$

where i is i^{th} subband, j is the frame count of the speech sample, b is the sample number, l is the frame length, and N is the frame shift duration.

NSD output, which is logarithmic in this case, is further applied to the scales of loudness functions [204]. Logarithmic operation also helps to reduce the dynamic range of data. Finally, DCT is applied to obtain the cepstral representation, subsequently normalized with Cepstral Mean Normalization (CMN) to obtain the static ETECC feature set. In order to extract *transitional* information, velocity (Δ) and acceleration ($\Delta\Delta$) coefficients are appended along with the static feature coefficients in order to form *static + dynamic* feature vector.

4.4.3 Spectrographic Analysis

4.4.3.1 TEO vs. ETEO Profiles

Figure 4.5 depicts TEO and ETEO profiles for a sample speech utterance found in ASVSpooF 2017 version-2 dataset. Specifically, Figure 4.5-(a) partially magnified in (b), and Figure 4.5-(c), partially magnified in (d), show TEO and ETEO profiles for 5^{th} and 15^{th} subband output signals, respectively. From Figure 4.5-(a), we can observe that both TEO and ETEO profiles are completely overlapping for lower subband signals. This is because, for lower frequencies, the assumption $\sin(\omega) \approx \omega$ holds true [182]. This means that the denominator in eq. (4.28) equals the unity, giving identical TEO and ETEO profiles. Furthermore, low frequency subband signals are smooth and slowly-varying. By observing eq. (4.32), it can be argued that the signal mass will be approaching to its maximum value for low frequency signals. This fact can be observed in Figure 4.5-(a) and Figure 4.5-(b), where TEO and ETEO profiles seem to be overlapping (almost exactly) for low frequency subband signals.

Contrary to this, for higher frequencies, Kaiser's approximation in eq. (4.11) is no longer valid and consequently, significant differences in TEO and ETEO profiles are observed suggesting that the latter gives an exact estimate for signal's energy. This fact can be observed in Figure 4.5-(c) and Figure 4.5-(d), which show the non-overlapping profiles for TEO and ETEO. Mathematically, when $\omega = 0$, i.e., $\text{sinc}^2(\omega) = 1$ then, ETEO profile will be similar to that of TEO. Increment in ω leads to reduction in $\text{sinc}^2(\omega)$ term and hence, signal mass also reduces. Ad-

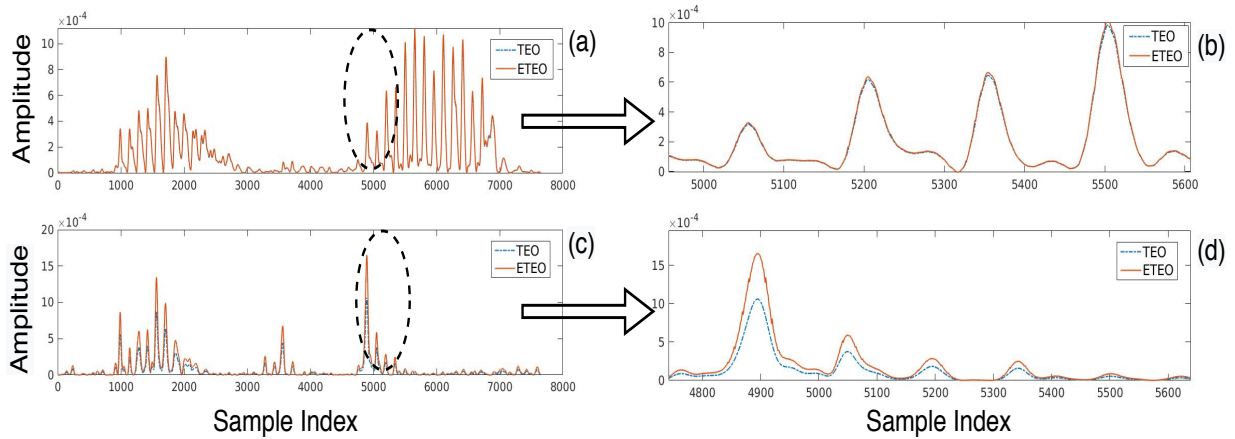


Figure 4.5: Energy Estimated by TEO and ETEO for (a) 5th and (c) 15th Subband Filter Output. A Magnified View for the Corresponding Encircled Region is Shown in (b), and (d), Respectively.

ditionally, from eq. (4.27) and eq. (4.28), we can infer that the energy estimated using ETEO is greater than that obtained by using TEO for the higher frequencies.

4.4.3.2 Waterfall Plot of TECC *vs.* ETECC Feature Sets

Figure 4.6 shows the waterfall plots for TECC and ETECC feature sets in three dimensions. It also demonstrates the capability of ETECC feature set against the TECC feature set for replay SSD task. Panel-I and Panel-II show the waterfall plots for the genuine speech and its corresponding spoofed speech chosen from ASVSpooof 2017 Version-2 dataset. Since replay speech signals can be modeled as being the convolution of the genuine speech signal with transmission channel effects induced by replay mechanism, which are bandpass in nature, spectral spread in the replay signal appears in the marked region with dotted square in the Panels. Additionally, for higher frequencies, log-energies estimated using ETEO are higher than those of TEO. The reasoning for this fact is discussed in Section 4.4.3.1. In the literature, it is also shown that the higher frequencies are more important for replay SSD, and since ETEO provides an exact estimate of signal’s energy for the higher frequencies as particularly discussed in Section 4.4.1. Furthermore, experimental results in subsequent sections shows that ETEO-based ETECC feature set shows relatively the best performance.

4.4.4 Experimental Setup

To evaluate the performance of the ETECC feature set, several datasets for anti-spoofing research, namely, ASVSpooof-2015, -2017, -2019, BTAS, and ReMASC are

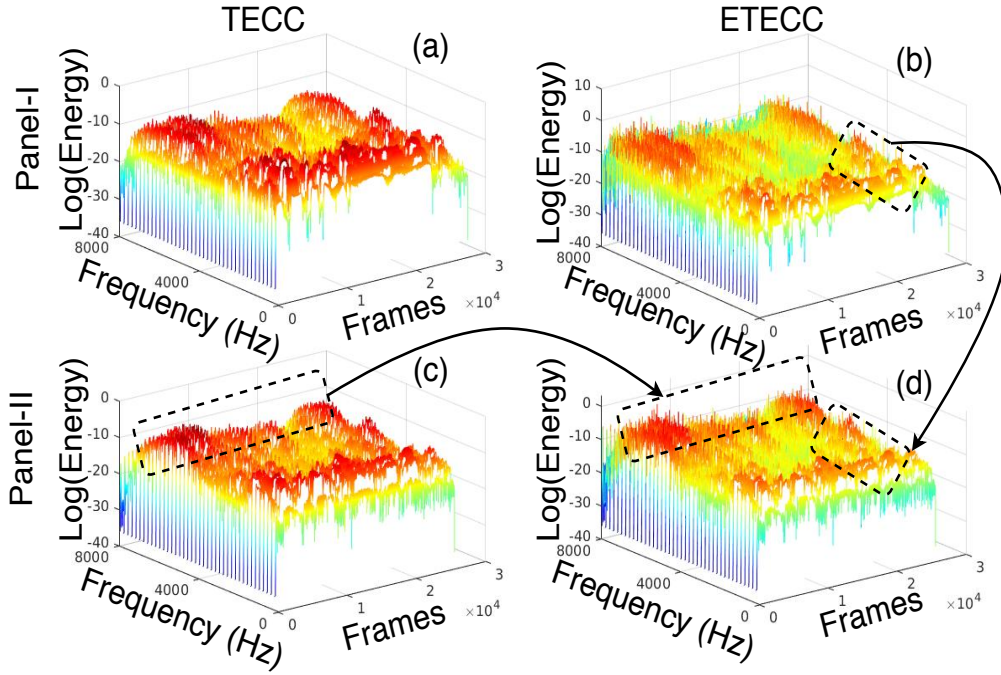


Figure 4.6: Waterfall Plot of *Genuine* (Panel I) and *Spoof* (Panel II) Speech Utterances: (a), (c) Waterfall Plots Obtained using TECC Feature Set, and (b),(d) Waterfall Plots Obtained using ETECC Feature Set. After [32].

utilized with corresponding dataset configuration shown in Table 3.1, Table 3.3, Table 3.5, Table 3.10, and Table 3.15, respectively. The experiments are also extended for environment-dependent scenario on ASVSpooof 2017 and ReMASC datasets, as discussed in Section 3.2.2 and Section 3.2.7, respectively.

The cepstral feature sets in this study are mainly categorized as spectral-based and subband-filtering-based features. The CQCC, MFCC, and LFCC feature sets are spectral-based feature sets, where magnitude in frequency-domain representation is utilized to estimate the cepstrum. On the other hand, TECC, SECC, and ETECC are processed through subband filters, where the impulse response of the subband filters are convolved with the speech signal. Even though all these six feature sets are cepstral feature, there is a slight difference in the way cepstrum is computed, either via frequency-domain subband or time-domain subband signals.

It has been observed that feature normalization techniques, such as CMN and Cepstral Mean and Variance Normalization (CMVN) are performing better on ASVSpooof 2017 version 2.0 dataset. Hence, these normalization techniques are applied on ASVSpooof 2017 version 2.0 dataset only. Furthermore, it is also observed that CMN works better for subband energy-based features and hence, the TECC, SECC, and ETECC feature sets are normalized by CMN. Other features in

our study, i.e., CQCC, LFCC, and MFCC are normalized by CMVN.

The study on ETECC feature set utilizes the GMM, CNN, and LCNN classifiers at the back-end. The parameter tuning for the GMM classifier is analyzed in Section 4.4.5.2. The CNN and LCNN models are implemented as described in Section 3.4. These models require constant input size. The proposed ETECC feature set is 120-dimensional (120-D) representation, and we kept the number of frames constant to 400 by appending or concatenating the required number of frames of a given utterance. Hence, feature representation of the proposed ETECC feature set becomes of size 120×400 for CNN and LCNN classifiers. For implementation of the CNN for 120-D feature set, we used Rectified Linear Unit (ReLU) as a non-linear activation function along with batch normalization. Xavier’s initialization was used for convolutional layers [205]. Adaptive moment estimation (ADAM) optimizer with momentum of 0.9 and a learning rate of 10^{-4} was used for training [206]. We used five convolutional layers followed by two fully connected (FC) layers. The last FC layer uses the softmax function to generate the scores, which is further utilized for evaluation of the system. For LCNN classifier, ReLU activation function is replaced by MFM activation. Other hyperparameters for LCNN classifier remains the same as that of CNN classifier. The details of the CNN and LCNN architecture for the 120-D feature representation is described in Table 4.1 and Table 4.2, respectively. For the other feature sets, the model architecture of CNN and LCNN classifiers is modified based upon the dimension of the feature vector. The CNN and LCNN models are trained upto 20 epochs. The training loss is computed for every epoch and the model was saved if it produces the least amount of % EER on the Dev set. The experiment is ran on the Eval set for the model, which shows the least amount of % EER on the Dev set.

For the study on ETECC feature set, EER and t-DCF are utilized as performance metrics, whereas the score-level fusion is performed using two popular approaches, namely, *linear weighted sum* and *logistic regression solution* [181]. The evaluation metrics and score-level fusion methods are described in Section 3.5 and Section 3.6, respectively.

4.4.5 Experimental Results

4.4.5.1 Paraconsistent Feature Engineering (PFE) for Ranking

In this work, paraconsistent logic (PL), by means of Paraconsistent Feature Engineering (PFE) is used to quantitatively analyze the efficacy of proposed ETECC feature set for replay SSD task on ASVSpooof 2017 version 2.0 dataset [207]. An

Table 4.1: Details of the Proposed CNN Architecture for SSD System. After [10].

Layer	Filter/Stride	Output	# Parameters
Conv1	5x5/1x1	16 x 120 x 400	416
MaxPool1	2x2/1x2	16 x 60 x 200	-
Conv2	3x3/1x1	32 x 60 x 200	4640
MaxPool2	2x2/1x2	32 x 30 x 100	-
Conv3	3x3/1x1	64 x 30 x 100	18496
MaxPool3	2x2/2x2	64 x 15 x 50	-
Conv4	3x3/1x1	16 x 15 x 50	9232
MaxPool4	2x2/2x2	16 x 7 x 25	-
Conv5	3x3/1x1	16 x 7 x 25	2320
MaxPool5	2x2/2x2	16 x 3 x 12	-
FC6	-	1 x 200	115400
FC7	-	1 x 2	402

ideal feature set for a certain classification task should exhibit intra-class similarities as well as inter-class distinction. Based on PFE, such an analysis is performed, initially by normalizing the feature sets in the range $0 \sim 1$. Among a variety of techniques for normalization, we performed the operation for which the sum of feature vector representation is the unity. For this reason, we used softmax function, which maps the D -dimensional input feature vector Z to the *pdf* consisting of D number of probabilities and thus, normalizing the features in the range $[0,1]$. This conversion of the feature vector to the *pdf* using softmax function not only ensures the intended normalization, but also forces the sum of the feature vector elements to be exactly 1 [174]. The softmax function $\rho(\cdot)$ is mathematically written as:

$$\rho(Z_i) = \frac{e^{Z_i}}{\sum_{j=1}^D e^{Z_j}}, \quad \text{for } i = 1, 2, \dots, D. \quad (4.39)$$

Once our normalized feature vectors are obtained, intra-class similarities and inter-class distinction are defined by the parameters α and β , respectively. To calculate α , the similarity vector for each dimension of the feature vector is computed. Let us consider 1 – *dimensional* feature representation for K number of samples, as $X[\cdot] = [x_1, x_2, \dots, x_K]$. Then, the deviation for the set of K real numbers will be the difference between the largest and smallest values in X , i.e.,

$$A = L(X[\cdot]) - S(X[\cdot]). \quad (4.40)$$

To find an advantageous intra-class similarities, A should be as small as pos-

Table 4.2: Details of the Proposed LCNN Architecture for SSD System. After [10].

Layer	Filter/Stride	Output	# Parameters
Conv1	5x5/1x1	32 x 120 x 400	832
MFM1	-	16 x 120 x 400	-
MaxPool1	2x2/1x2	16 x 60 x 200	-
Conv2a	1x1/1x1	32 x 60 x 200	544
MFM2a	-	16 x 60 x 200	-
Conv2b	3x3/1x1	64 x 60 x 200	9280
MFM2b	-	32 x 60 x 200	-
MaxPool2	2x2/1x2	32 x 30 x 100	-
Conv3a	1x1/1x1	64 x 30 x 100	2112
MFM3a	-	32 x 30 x 100	-
Conv3b	3x3/1x1	128 x 30 x 100	36992
MFM3b	-	64 x 30 x 100	-
MaxPool3	2x2/2x2	64 x 15 x 50	-
Conv4a	1x1/1x1	128 x 15 x 50	8320
MFM4a	-	64 x 15 x 50	-
Conv4b	3x3/1x1	64 x 15 x 50	36928
MFM4b	-	32 x 15 x 50	-
MaxPool4	2x2/2x2	32 x 7 x 25	-
Conv5a	1x1/1x1	64 x 7 x 25	2112
MFM5a	-	32 x 7 x 25	-
Conv5b	3x3/1x1	32 x 7 x 25	9248
MFM5b	-	16 x 7 x 25	-
MaxPool5	2x2/2x2	16 x 3 x 12	-
FC6	-	1 x 128	73856
MFM6	-	1 x 64	-
FC7	-	1 x 2	130

sible. Let $Y = 1 - A$ be the standard similarity measure, where $Y \approx 0$ indicates low similarity, and $Y \approx 1$ indicates high similarity between feature vectors of the same class. For multi-dimensional feature representation, the computation of Y for each dimension should be carried out independently to form a similarity vector, which is computed for each class. Let us consider that the given classification task consist of N number of classes and similarity vector for class- c can be represented as $Y_c = [y_c(1), y_c(2), \dots, y_c(D)]$, being D the dimension of the feature

vector. The intra-class similarity for class- c is computed as in [207], i.e.,

$$\bar{Y}_c = \frac{1}{D} \sum_{i=1}^D y_c(i). \quad (4.41)$$

To assess the worst-case similarity among all the classes, α is defined as:

$$\alpha = \min\{\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_N\}. \quad (4.42)$$

The inter-class distinction is estimated by computing the two range vectors for each class of size, D . One of the range vector consists of minimum value computed over each dimension for feature vector. Other range vector gives the maximum value obtained over each dimension for the entire feature set. This computation is nicely depicted in [207]. These range vectors are expected to store the interval in which all the feature vectors of a class should lie. Once range vectors are obtained, dimension-wise comparison between each feature vector of one class and the range vector of all the other classes is performed to determine the number of overlapping feature vector elements, Z .

Each overlap indicates that a feature vector from one class conquered the range of the other class, which is undesirable for the classification task. Next, β is defined as [207]:

$$\beta = \frac{Z}{F}, \quad (4.43)$$

where F is the maximum possible number of overlaps. Considering that each of the N classes accommodate S feature vectors of size D , then we can easily find that $F = N \cdot (N - 1) \cdot S \cdot D$.

The parameters α and β can now serve to compute the degree of certainty, i.e., $G1 = \alpha - \beta$, and degree of contradiction, i.e., $G2 = \alpha + \beta - 1$. Considering that $0 \leq \alpha, \beta \leq 1$, then $-1 \leq G1, G2 \leq 1$. The two-dimensional plane, where $G1$ and $G2$ are defined, is known as paraconsistent plane. For better performance of the feature sets, the point $P = (G1, G2)$ should lie nearer to the corner $(1, 0)$, as it is the ideal case for the feature set to be fully linearly-separable. The distance between the point $P = (G1, G2)$ for each feature set and the reference point $Q = (1, 0)$ in the paraconsistent plane is denoted as $d(P, Q)$ and is estimated by using the Euclidean distance formula. The overall procedure of PFE is briefly illustrated in Algorithm 2 [207].

In our application, experiments are performed on the training set of ASVSpooof 2017 version-2 dataset for paraconsistent feature analysis. The training data consists of 1507 utterances for each of the genuine as well as spoof speech class [3].

Algorithm 2 MATLAB Pseudo Code to implement the PFE. After [207].

1. Normalize the feature set using softmax function,
 2. Intra-class Similarity measurement:

for $c = 1 : N$ do ,	N - number of classes,
for $i = 1 : D$ do ,	D - number of dimensions of feature vector,
$A_c(i) = L(X_c(i,:)) - S(X_c(i,:))$,	compute deviation,
$Y_c(i) = 1 - A_c(i)$,	similarity index,
end for	
$\bar{Y}_c = \frac{1}{D} \sum_{i=1}^D y_c(i)$,	arithmetic mean along all dimensions,
end for	
$\alpha = \min\{\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_N\}$,	worst case intraclass similarity,
 3. Inter-class Dissimilarity measurement:

for $c = 1 : N$ do ,	N - number of classes,
for $i = 1 : D$ do ,	D - number of dimensions of feature vector,
$V_{min}(i) = \min(X(i,:))$,	elementwise minimum values of whole set,
$V_{max}(i) = \max(X(i,:))$,	elementwise maximum values of whole set,
end for	
Compute Z ,	Z - number of overlapping feature elements,
end for	
$F = N \cdot (N - 1) \cdot S \cdot T$,	S - number of feature vectors
$\beta = \frac{Z}{F}$	
 4. $G1 = \alpha - \beta$, degree of certainty,
 5. $G2 = \alpha + \beta - 1$, degree of contradiction,
 6. $P = (G1, G2)$, point in paraconsistent plane,
 7. Find distant $d(P, Q)$, where $Q = (1, 0)$.
-

The results obtained by using paraconsistent analysis over CQCC, LFCC, MFCC, cepstral, SECC, TECC, and ETECC feature sets are shown in Table 4.3. Particularly $d(P, Q)$ for both TECC and ETECC are the smallest, i.e., they are linearly-separable due to high intra-class similarity and low inter-class dissimilarity and hence, they are better feature sets for the given classification task.

4.4.5.2 Results on ASVSpooof 2017 Version-2 Dataset

- **Effect of the Type of Filterbank**

The proposed feature set adopts a Gabor filterbank using linearly-spaced sub-band filters. The performance of the proposed feature set with Gabor filterbank was compared with the other filterbanks, namely, Gammatone, Cochlear, and Mexican hat, which are used in the speech literature. Gammatone filterbank describes the shape of the impulse response function of the auditory system as es-

Table 4.3: Evaluation of Various Feature Sets with Paraconsistent Framework on ASVSpooof 2017 Version-2 Dataset. After [10].

Feature Set	α	β	G1	G2	$d(P, Q)$
CQCC	0.48	1	-0.52	0.48	1.5940
LFCC	0.55	0.99	-0.44	0.54	1.5379
MFCC	0.29	1	-0.71	0.29	1.7344
Cepstral	0.48	0.99	-0.51	0.47	1.5814
SECC	0.65	0.99	-0.34	0.64	1.4850
TECC	0.66	0.99	-0.33	0.65	1.4803
ETECC	0.66	0.99	-0.33	0.65	1.4803

timated by the reverse correlation function of neural firing times [208, 209]. The Gammatone filter is defined in the time-domain (impulse response function) as:

$$gt(t) = t^{n-1} \cdot \exp(-2\pi bt) \cdot \cos(2\pi f_c t + \theta) \cdot U(t), \quad (4.44)$$

where n controls the relative shape of the envelope, b controls the duration of the impulse response function, f_c determines the frequency of the carrier, θ represents the phase of the carrier, and $U(\cdot)$ represents the unit-step function. All the four parameters have corresponding effects on the frequency-domain characteristics of the Gammatone filter. The eq. (4.44) represents an amplitude modulated carrier tone of frequency f_c with an envelope proportional to $t^{n-1} \cdot \exp(-2\pi bt)$, which is the familiar with Gamma distribution from statistics. Considering the nature of these two components, the filterbank is known as Gammatone filterbank. The Mexican hat filterbank is derived from the negative normalized second-order derivative of a Gaussian function [210]. The details of the cochlear filterbank can be studied in Section 4.6.1. Furthermore, the magnitude of the frequency responses of the various filterbanks utilized in this study, are depicted in Figure 4.7 to understand their statistical behavior. Notably, from Figure 4.7, it can be observed that:

- variance of the frequency responses of subband filters in frequency-domain is constant for Gabor and Mexican-hat filterbank.
- variance of the frequency response of subband filters in frequency-domain is increasing with increase in frequency. It is very high for higher frequency subband filters.

Additionally, authors also analyzed the spectral energy densities where energies are estimated using ETEO (abbreviated as ETEO-spectrogram). ETEO-

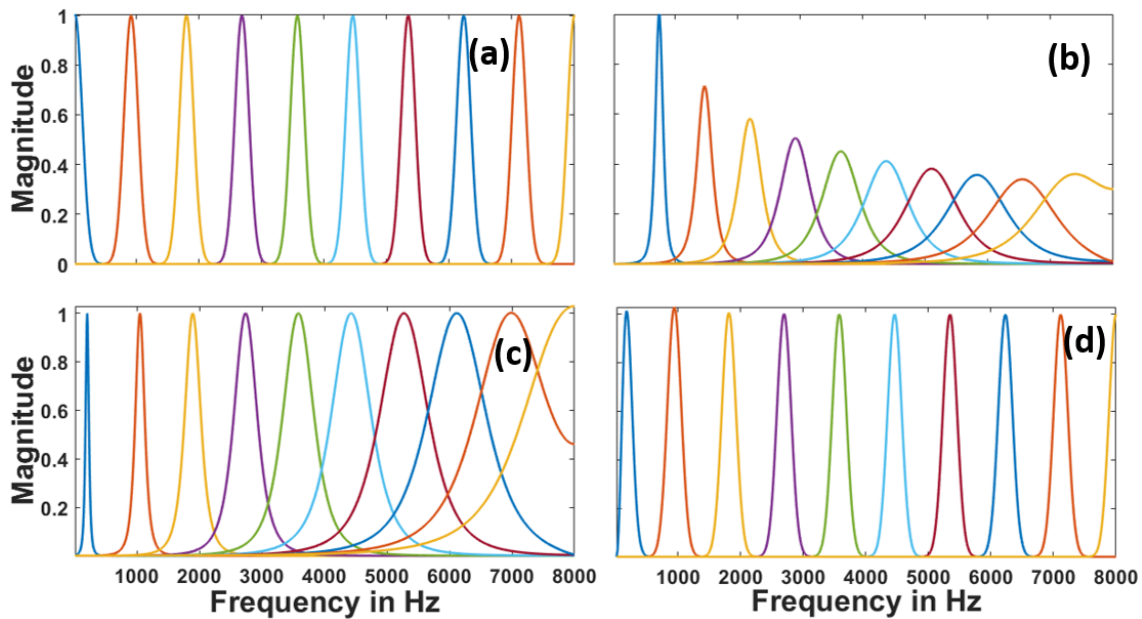


Figure 4.7: Frequency Response of (a) Gabor Filterbank, (b) Cochlear Filterbank, (c) Gammatone Filterbank, and (d) Mexican-hat Filterbank.

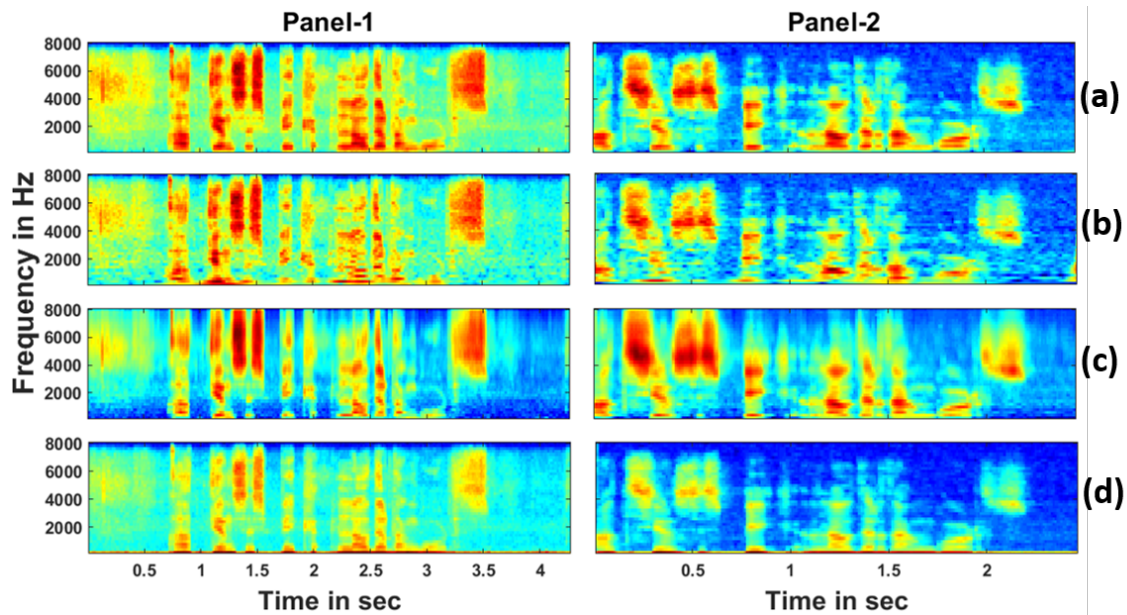


Figure 4.8: ETEO-spectrogram Representation Obtained from (a) Gabor, (b) Cochlear, (c) Gammatone, and (d) Mexican-hat Filterbank with 40 Subband Filters in the Filterbank for Genuine (Panel-1), and it's Corresponding Spoof (Panel-2) Speech Utterance.

Table 4.4: Results (in % EER) for ETECC Feature Set *w.r.t.* Type of the Filterbank. After [10].

Filterbank Type	Dev	Eval
Gammatone	29	30.21
Mexican hat	9.11	29.77
Cochlear	11.82	16.56
Gabor	5.55	10.75

spectrogram is obtained for three randomly selected genuine and their corresponding spoof speech utterances using various filterbanks utilized in this study. Spectrograms for one of the genuine *vs.* spoof utterance are depicted in Figure 4.8, where it can be observed that:

- Gabor filterbank has *optimal* time-frequency resolution. This means that Gabor filter tend to emphasize approximately equally in both the domains. This fact, however, is not true for the other filterbanks included here for analysis. For instance, observing Figure 4.8(a) and Figure 4.8(b), it can be easily said that Gabor filterbank gives a more profound representation than the other filterbanks for both the genuine and spoof speech utterances.
- when comparing Figure 4.8(a) and Figure 4.8(c), we can say that Gammatone filterbank has comparatively much poor representation both in lower and high frequency regions. Smearing in spectrograms is observed for Gammatone filterbank, which is somehow related to the time-domain representations of both the subband filters. The similar observations can be noticed for the other pair of utterances.

The center frequencies of the subband filters in this filterbank are also linearly-spaced. The performance of ETECC feature set on varying the type of the filterbank was compared with % EER metric, as shown in Table 4.4. The best performance is observed for the feature set extracted using a Gabor filterbank. This may be due to the fact that Gabor filters are smooth, possess excellent time-frequency resolution and maintain their shape across entire frequency range (discussed in sub-Section 4.4.2).

- **Parameter tuning**

Initially, training and Dev subsets were used to finetune the feature and model parameters, such as number of subband filters in the Gabor filterbank, bandwidth of the subband filters, frame length, dimension of feature vector, the number of

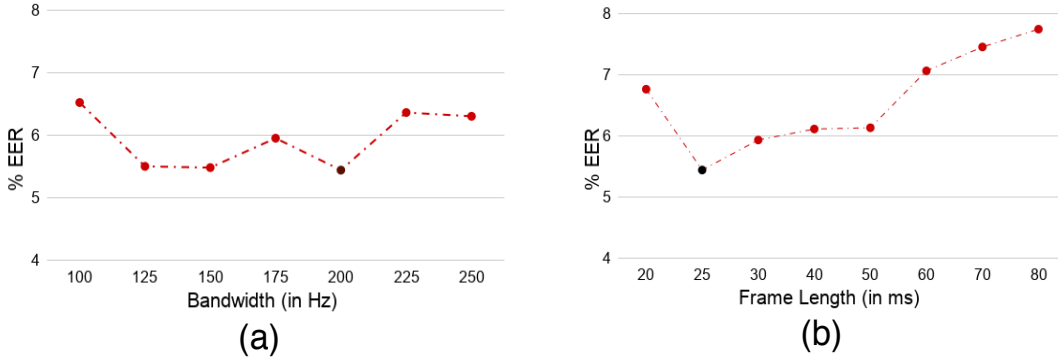


Figure 4.9: Variation of % EER of Dev set with the (a) Bandwidth of Subband Filters in the Gabor Filterbank, and (b) Speech Frame Length of Analysis Window during ETECC Feature Extraction. After [10].

mixtures in GMM, frequency region of subband filtered representation, and static *vs.* dynamic features. The system performance was then assessed over the Eval subset with the parameters previously tuned on the Dev set.

- *Effect of Bandwidth of the Subband Filters*

The authors of paper [197] showed that the quality of the subband features extracted for SSD task depends on their half-power bandwidth. Hence, the effective Root Mean Square (RMS) bandwidth of bandpass filters is varied from 100 Hz to 250 Hz. For these experiments, frame length is set to 25 ms as vocal tract system is reasonably modeled as a Linear Time-Invariant (LTI) system within this short duration of frame length. In the previous studies of TECC and ETECC feature extraction, the number of subband filters and the number of cepstral coefficients were set to 40 and 120, respectively, to provide the relatively best SSD performance [32, 106]. Hence, we initialized with the similar setting for these two parameters. Results obtained by varying the bandwidth are shown in Figure 4.9-(a). Clearly, the best performance was obtained by tuning the bandwidth to 200 Hz and hence, the center frequency of subband filter was fixed to that frequency for the next set of experiments.

- *Effect of Frame Length for Analysis Window*

Experiments were also performed to fine-tune the appropriate value of the frame length by varying it from 20 ms to 80 ms. Meanwhile, the bandwidth of each subband filter was set to 200 Hz as discussed earlier, whereas all the other feature and model parameters were kept the same. From Figure 4.9(b), it can be observed that a lower % EER was obtained for a frame length

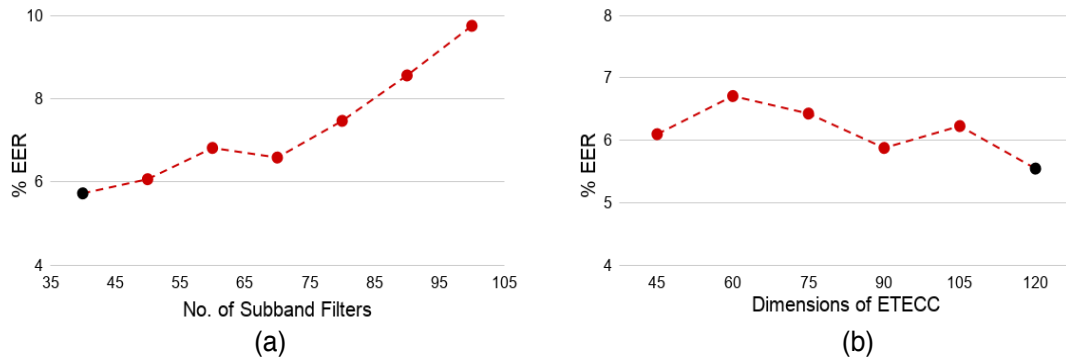


Figure 4.10: Variation of % EER of Dev Set with the (a) Number of Subband Filters, and (b) Dimensions of ETECC Feature Vector (Including Static, Δ , and $\Delta\Delta$ Coefficients). After [10].

of 25 ms for analysis window. Beyond this point, EER increases as the frame length increases and hence, a frame length of 25 ms was kept for the remaining set of experiments.

- *Effect of Subband Filters in Gabor Filterbank*

Original investigations reported in [40] reveal that the human auditory system consists of thousands of auditory filters. This motivated us to analyze the effect of varying the number of subband filters in a Gabor filterbank. Results with varying the number of subband filters were reported in Figure 4.10(a). For this experiment, bandwidth and frame length were set to 200 Hz and 25 ms, respectively, as they provide the lowest % EER. Notably, the best performance on the Dev set was obtained by using 40 subband filters in the filterbank. We obtained 5.55 % EER on the Dev set. Figure 4.10(a) also suggests that the performance deteriorates consistently on the Dev set as the number of subband filters in the filterbank crosses 40. These results are in contradiction to our motivation suggested in [40]. One possible explanation might be that, as the number of subband filters increases, overlaps between any two consecutive subband filters occur in the frequency-domain and hence, discriminative information is lost, resulting in performance degradation for the SSD task.

- *Effect of Dimension of Feature Vector*

The experiments were also performed to fine-tune the optimal number of dimensions of the proposed ETECC feature set. ETECC feature vector contain static, Δ , and $\Delta\Delta$ coefficients. In this experiment, the dimension varies from 45 to 120, and the results are shown in Figure 4.10(b), which clearly reveals

Table 4.5: Results (in % EER) for Varying Number of Mixtures in GMM for the Features Extracted using 40 Subband Filters in Gabor Filterbank. After [10].

# Mixtures in GMM	Dev	Eval
64	5.77	11.83
128	5.51	11.88
256	5.82	11.06
512	5.55	10.75
1024	5.32	11.38

that 120-D ETECC feature representation offers relatively best performance for the replay SSD task. Results are better for higher dimensions (here 120) primarily due to reduction in *feature occupancy* and thereby increasing feature discrimination and inter-class separability in higher-dimensional feature space.

- *Effect of Number of Mixtures in GMM*

The number of mixtures to be used for training the GMM depends upon the amount of training data and dimension of the feature representation. If the size of the training data is small, with lesser dimension of feature representation, then a small number of mixtures are required to produce good results. As the size of the training data and dimension of the feature vector increases, the number of mixtures should also increase in order to model the overall data distribution properly. To select the optimum number of mixtures, experiments were performed with 120-D ETECC feature vector and, observably, better performance on the Eval set was obtained by using 512 mixtures in the GMM as shown in Table 4.5. This result indicate that a small number of mixtures (in particular, 64, 128, and 256) are not able to model the overall distribution properly, whereas degradation in results for 1024 mixtures indicate overfitting of the data using GMM.

- *Effect of Low, Mid, and High Frequency Subband Filters*

The authors of paper [195] found that the effect of replay mechanism contributes more in high frequency regions. To analyze such an effect, experiments were performed by selecting the low, mid, and high frequency regions and by choosing the corresponding subband filters in the filterbank to keep the remaining feature extraction scheme intact. The analysis was performed with 40 subband filters in the filterbank, as it produces relatively optimal results, as shown in Table 4.6. The prominent observation is that all the frequency regions are important

Table 4.6: Results (in % EER) for Subband Filters in Low, Mid, and High Frequency Regions. After [10].

Frequency Regions	Range of Subband Filters	ETECC		TECC	
		Dev	Eval	Dev	Eval
Low	1-20	21.94	23.67	22.12	23.51
Mid	11-30	35.75	29.97	34.72	29.56
High	16-40	12.12	17.15	12.78	17.60
High	21-40	12.50	18.39	11.99	18.75
Mid & High	11-40	10.44	15.06	10.69	15.97
All	1-40	5.55	10.75	5.87	11.34

to produce the better performance. However, high frequency regions contribute relatively more in the replay SSD task, as they cause significant reduction in % EER in comparison with the low and mid frequency subband filters in the filterbank. Thus, ETECC provides better results, albeit marginally, than the TECC for high frequency regions. This indicates the capability of ETEO to capture important details, via signal mass, specially in the high frequency regions, as illustrated in Section 4.4.1. However, it should be noted from Table 4.6 that TECC produces marginally better results than the ETECC for low and mid frequency regions.

- *Effect of Static vs. Dynamic Features*

This experiment was performed to analyze the individual contribution of the static, Δ , and $\Delta\Delta$ features of ETECC. Along with individual performance of these features, we also analyzed the performance of their possible combination. From Table 4.7, we can note that the performance of the Δ features is better than the static and $\Delta\Delta$ features alone. The combination of the static and Δ features gave % EER of 11.52 % on the Eval set, whereas the combination of static, Δ , and $\Delta\Delta$ features provide a % EER of just 10.75 %. Thus, the static and Δ features contribute more in producing the better results.

- **Results using Various Feature Sets and Classifiers**

The performance of CQCC (baseline), Cepstral, MFCC, TECC, LFCC, SECC, and ETECC feature sets is shown in Table 4.8 in terms of % EER on Dev and Eval sets. GMM is used as the back-end classifier for all of these feature sets. CMVN is used for CQCC, MFCC, Cepstral, and LFCC feature sets while CMN is used for TECC, SECC, and ETECC feature sets. The 90-D CQCC baseline provided 12.270 % and 18.810 % over Dev and Eval sets, respectively. On the other hand, 39-D state-of-the-art MFCC feature set provided 22.39 % and 25.34 % on Dev and Eval

Table 4.7: Results (in % EER) for Static, Δ , and $\Delta\Delta$ Features for ETECC Feature Set. After [10].

# Static vs. Dynamic Features	Dev	Eval
static (40-D)	7.02	16.09
Δ (40-D)	6.37	14.33
$\Delta\Delta$ (40-D)	9.17	17.46
static + Δ (80-D)	5.49	11.52
static + $\Delta\Delta$ (80-D)	6.92	13.41
Δ + $\Delta\Delta$ (80-D)	6.44	14.30
static + Δ + $\Delta\Delta$ (120-D)	5.55	10.75

sets, respectively. Also, the cepstral feature set derived from STFT shows the EER of 10.18 % and 15.30 % for Dev and Eval set, respectively. The performance of LFCC, SECC, and TECC feature sets with 120 elements was also evaluated for comparison. Furthermore, we extended the experiments with two deep learning classifiers, namely, CNN and LCNN.

ETECC-GMM SSD system produced 5.55 % and 10.75 % EER on Dev and Eval sets, respectively, whereas SECC-GMM SSD system produced 7.46 % and 13.84 % EER on Dev and Eval sets, respectively. Those results directly suggest that energy operator-based feature set performs well for replay SSD task. Also compared with alternate energy measures, such as TEO and squared energy function, ETECC feature set showed the best performance. An important absolute reduction of 6.72 % and 8.06 % in EER was observed, when the performance of ETECC feature set was compared to the baseline CQCC feature set. Besides this, we can note from Table 4.8 that ETECC performs better using deep learning architectures, i.e., CNN and LCNN over TECC feature set. Furthermore, it can be observed that the conventional GMM classifier is performing better than the deep learning-based CNN and LCNN classifiers. It can be due to the data being relatively better approximated to be Gaussians and the characteristics are better suited for GMM. This kind of results are also observed for further experimental results obtained in this thesis. Notably, in the ASVspoof 2021 PA challenge, the LFCC-GMM baseline (with 39.79 % ERR) showed better performance as compared to LFCC-LCNN baseline (with 42.16 % ERR) [170]. This also shows that GMM can indeed perform better than the CNN.

Interestingly, score-level fusion of our ETECC feature set with the other feature sets was carried out by using two methods, and the results are reported in Table 4.9. The first method of fusion is linear regression, where two scores were combined as per eq. (3.16). Fusion results of ETECC with all the other feature sets

Table 4.8: Results (in % EER) on Eval and Dev Datasets for Individual SSD Systems. After [10].

Feature Sets	% EER	
	Dev	Eval
CQCC-GMM (Baseline) (S1)	12.27	18.81
MFCC-GMM (S2)	22.39	25.34
Cepstral-GMM (S3)	10.18	15.30
LFCC-GMM (S4)	14.44	13.73
TECC-GMM (S5)	5.87	11.34
SECC-GMM (S6)	7.46	13.84
ETECC-GMM (S7)	5.55	10.75
TECC-CNN (S8)	8.35	14.33
TECC-LCNN (S9)	7.19	14.05
ETECC-CNN (S10)	8.83	15.02
ETECC-LCNN (S11)	6.98	13.54

were found to be approximately the same.

Table 4.9: Results (in % EER) on Score-Level Fusion on Eval and Dev Datasets using Linear and Logistic Regression. After [10].

Score-Level Fusion of SSD Systems	Linear Regression		Linear Logistic Regression (using Bosaris Toolkit) [211]	
	Dev	Eval	Dev	Eval
S1 + S7	5.55	10.54	5.75	10.59
S2 + S7	5.14	10.53	5.56	11.24
S3 + S7	5.55	10.68	4.75	10.81
S4 + S7	5.55	10.67	5.17	10.83
S5 + S7	5.40	10.69	5.51	10.87
S6 + S7	5.51	10.69	5.65	10.74
S1 + S2 + S3 + S4 + S5 + S6	5.33	10.62	6.78	11.31
S1 + S2 + S3 + S4 + S5 + S6 + S7	4.69	10.43	5.61	11.27
S7 + S10 + S11	4.91	10.46	6.64	12.67

System S1 - S10 are as per Table 4.8. '+' indicates the score-level fusion of systems.

The performance of the above mentioned feature sets can, in addition, be inspected based on a DET curve [179]. Observing the DET plots, as shown in Figure 4.11, we can see that the miss probabilities of CQCC and LFCC feature sets are notably high for any given value of false alarm probability. This is an unwanted characteristic for a good SSD system.

The LLR score distribution of genuine *vs.* impostor utterances on Eval and

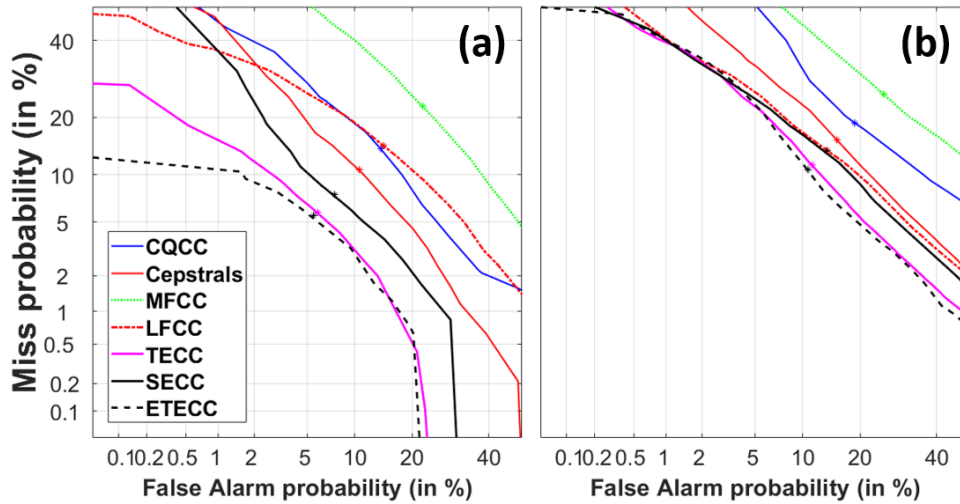


Figure 4.11: Individual DET Curves on (a) Dev, and (b) Eval Sets. After [10].

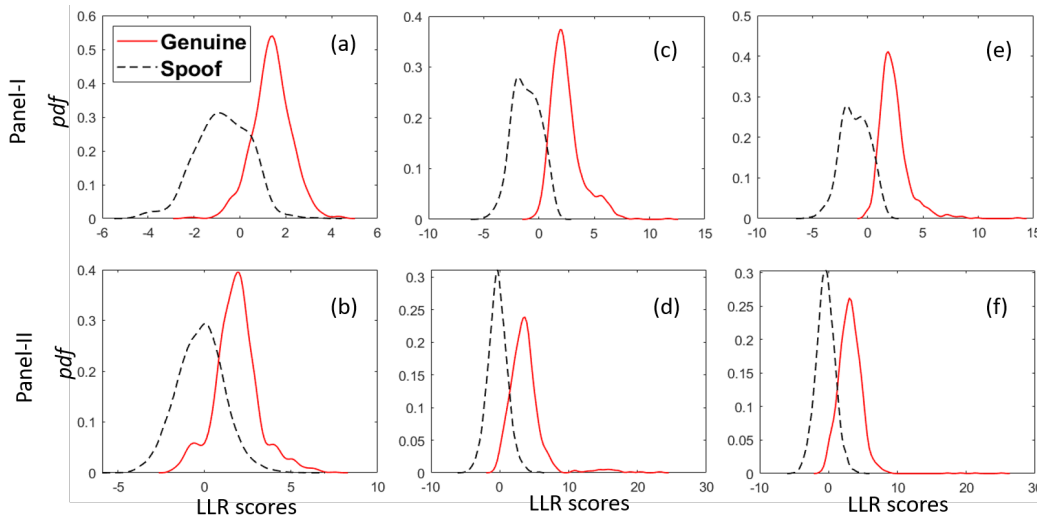


Figure 4.12: LLR Scores Distribution on Dev Set, (Panel I) and Eval Set (Panel II) using (a),(b) CQCC, (c),(d) TECC, and (e),(f) ETECC Feature Sets. After [10].

Dev sets are shown in Figure 4.12. The common area under the two Gaussian-like curves in Figure 4.12 is just the probability of misclassification [212]. We call this area as intersecting area. The lesser the intersecting area is, the lesser the probability of misclassification is and hence, the lesser the % EER for such a feature set will be. Visually, we observe that the intersecting area is relatively smaller for ETECC feature set, followed by TECC feature set and then the baseline CQCC feature set. This is an indication of a better classification (in particular, feature discrimination) ability of the proposed ETECC feature set.

- **Results for the Environment-Dependent Scenario**

Table 4.10: Results (in % EER) of Environment-Dependent Case on ASVSpooof 2017 Dataset. After [10].

Environment	CQCC	TECC	ETECC
Anechoic Room	0.26	0.20	0
Analog wire	11.42	11.78	9.10
Balcony	0	0	0
Canteen	0.93	0.17	0.46
Home	2.12	2.12	1.54
Office	5.63	4.37	2.35
Studio	0	0	0

The experiments performed for the environment-dependent scenario are reported in Table 4.10. In this case, the individual system was trained and tested for single environment only. The distribution of the number of spoof speech utterances for each environment is varying and shown in Table 3.4. To develop an individual environment-dependent SSD system, half of the spoofed speech utterances for the corresponding environment were chosen for training, whereas the remaining half were used for testing the model performance. To train the genuine and the spoofed speech models, an equal number of utterances were selected. In particular, environment-dependent-based SSD performs much better for un-normalized (i.e., without CMN/CMVN) feature sets. For the standard protocols provided by ASVSpooof 2017 challenge organizers, Eval set consists of speech samples for seen and unseen environments [3]. For these protocols, normalized feature sets perform better than the un-normalized ones. This contradictory behavior was analyzed in [21] for the baseline CQCC feature set. Additionally, ETECC feature set performs better for all the environments over the CQCC feature set. Furthermore, for three environments, namely, anechoic room, balcony, and studio, it provides 0 % EER.

4.4.5.3 Results on ASVSpooof 2019 Challenge, BTAS, and AVSspooof 2015 Challenge Datasets.

Following the same details mentioned above for the previous procedures, the results on ASVSpooof 2019 challenge for PA scenario, ASVSpooof 2019 challenge for LA scenario, BTAS, and AVSspooof 2015 challenge datasets can be found in Tables 4.11, 4.12, 4.13, and 4.14, respectively. It can be observed from Table 4.13, that the proposed ETECC feature set performs better than the TECC and baseline CQCC feature sets for BTAS dataset. In addition, proposed ETECC feature sets performs

better than the CQCC feature set for LA-scenario in ASVSpooof 2019 challenge dataset as shown in Table 4.12. However, TECC performs slightly better than the ETECC feature set. Furthermore, ETECC feature set could not perform better than the baseline CQCC feature set for PA scenario in ASVSpooof 2019 challenge dataset and ASVSpooof 2015 challenge dataset (Table 4.11 and Table 4.14). This could be the limitation of our proposed feature set that it could not perform better for all the datasets designed for anti-spoofing research.

4.4.5.4 Results on ReMASC Dataset

- Results on Dev and Eval Sets

To check the efficacy of ETECC feature set over ReMASC, experiments were performed according to the dataset configuration given in Table 3.15. The comparison was carried out with the other conventional feature sets for replay SSD task, namely, CQCC, LFCC, and MFCC. The experiments were also performed with TECC feature set so that we can validate the efficiency of energy tracking with the novel signal mass concept embedded in the proposed ETECC feature set. The results are shown in Table 4.15. The absolute reduction in % EER of 4.12 % and 7.89 % was observed for ETECC feature set against the baseline CQCC-GMM SSD system, for Dev and Eval subsets, respectively. Furthermore, the absolute reduction in % EER of 2.12 % and 1.19 % was observed for ETECC feature set against the TECC, for Dev and Eval subsets, respectively.

- Results on Environment-Dependent *vs.* Independent Scenarios

Experiments were also performed for environment-dependent *vs.* independent scenarios. For the former, the target environment is already seen by the defense model, whereas for environment-independent scenario, the defense model was trained on any of the three environments and tested on the fourth environment. This set of experiments are useful to compare and analyze the effect of environment-independent *vs.* -dependent scenario, in particular, how independence of the environment contribute to the difficulty in the SSD task. In addition, the environment-dependent scenario is important for realistic applications of VAs. In particular, VAs are primarily used in home applications and hence, the acoustic environment for training and testing is expected to be similar, if not identical.

For the experiments on environment-dependent scenario, each environment was partitioned into two disjoint and speaker-independent sets of roughly the same size. The environment-wise statistics of the ReMASC dataset are shown in

Table 4.11: Results (in % EER) on ASVSpooof 2019 Challenge Dataset for PA Scenario. After [10].

SSD System	Dev		Eval	
	t-DCF	EER	t-DCF	EER
CQCC-GMM (<i>Baseline</i>)	0.2007	10.25	0.2499	11.44
TECC-GMM	0.2113	9.58	0.2815	11.60
ETECC-GMM	0.2166	9.85	0.2845	11.77

Table 4.12: Results (in % EER) on ASVSpooof 2019 Challenge Dataset for LA Scenario. After [10].

SSD System	Dev		Eval	
	t-DCF	EER	t-DCF	EER
CQCC-GMM (<i>Baseline</i>)	0.0145	0.4709	0.2687	11.10
TECC-GMM	0	0	0.1709	6.87
ETECC-GMM	0	0.0022	0.1718	7.36

Table 4.13: Results (in % EER) on BTAS Dataset. After [10].

SSD System	Dev	Eval
CQCC-GMM (<i>Baseline</i>)	2.57	4.45
TECC-GMM	2.13	4.99
ETECC-GMM	1.50	2.95

Table 4.14: Results (in % EER) on ASVSpooof 2015 Challenge Dataset. After [10].

SSD System	Dev	Eval
	EER	EER
CQCC-GMM (<i>Baseline</i>)	0.03	4.57
TECC-GMM	0.13	7.80
ETECC-GMM	0.14	8.15

Table 4.15: Results (in % EER) on ReMASC Dataset using GMM Classifier. After [10].

SSD System	Dev	Eval
CQCC (<i>Baseline</i>)	20.57	23.31
LFCC	28.89	26.31
MFCC	36.43	31.53
TECC	18.57	16.61
ETECC	16.45	15.42

Table 3.14. The results obtained using CQCC, TECC, and ETECC feature sets with GMM classifier are reported in Table 4.16. Particularly for this scenario, we reported the results with the application of the CMVN for each utterance, as it has shown significant improvement. The analysis for the application of the CMN/CMVN techniques on environment-dependent *vs.* -independent scenario is discussed in Chapter 6 (Section 6.2). This needs further investigation and remains an open research problem. Notably, TECC performs better than the ETECC feature set for all the environments.

For an environment-independent scenario, results are shown in Table 4.16. It clearly indicates that the proposed ETECC feature set performs better than the TECC and CQCC feature sets, for three unseen environments, namely, environment A, B, and C, whereas all of the feature sets show the poor performance on environment D indicating environment D could not be expressed as linear combination of the other environments. In this scenario, results for environment-A perform well for un-normalized feature sets. For B and C as test environments, CMVN applied on feature set showed a significant improvement on the performance over un-normalized feature sets.

It can be also observed that TECC feature set performs better than the ETECC feature set in Env-D for environment-independent case. It might be due to noise suppression capability of the TEO especially for the vehicle noise (as originally reported for noise robust speech recognition in car [213]), whereas noise suppression capability of ETEO remains an open research question for the future study.

The performance of the ETECC feature set is summarized in Chapter Summary (Section 4.7). In the next sub-Section, the development and the performance of the proposed CTECC_{max} feature set is discussed.

Table 4.16: Results (in % EER) for Environment-Dependent *vs.* -Independent Case on ReMASC Dataset on GMM Classifier. After [10].

Acoustic Environment	Feature Set	Env-A	Env-B	Env-C	Env-D
Env-Dependent	CQCC	23.27	42.62	12.96	15.85
	ETECC	15.81	37.11	15.11	12.30
	TECC	15.09	33.77	14.11	10.35
Env-Independent	CQCC	35.65	40.89	35.95	49.99
	ETECC	28.99	32.45	29.97	49.91
	TECC	34.85	33.79	31.62	49.12
Env = Environment					

4.5 CTECC_{max} Feature Set

Modern VA technology is highly convenient to control the various household devices and applications. However, these devices are highly vulnerable to various spoofing attacks such as replay, impersonation, SS, and VC [48, 214]. Although, the design of CM systems for ASV and VAs looks similar, there are important differences in developing the anti-spoofing strategies for ASV and VAs. In particular, VAs are designed for long-distant speech recognition, where close-distant features could not be employed for the spoofed SSD task. In addition, ASV is generally implemented with a single microphone, whereas VAs generally make use of microphone array [9]. Considering these differences, ReMASC dataset has been released, which is specifically designed to develop countermeasures against replay spoofing attack on VAs [9]. In this work, CTEO-based feature set is proposed to build CM system against replay spoofing attacks for VAs.

The CTEO estimates interaction between two signals in terms of relative energy, as proposed in [215], in particular, CTEO shows the relative changes between two signals [216]. Furthermore, the link between TEO and ambiguity function was studied in [217], which helps to estimate the second moment angular bandwidth and the moments of a signal duration (spread), as well as that of its spectrum. CTEO for complex-valued signal was proposed in [218]. In [219], a novel similarity measure based on TEO was introduced for time series analysis, and the performance of this similarity measure is compared against the conventional approaches, namely, Euclidean distance and correlation coefficient. The CTEO was also used to redefine the general wave equations [220]. Furthermore, in [221], quadratic superposition law was used for transient detection. Other applications of the CTEO, such as medical imaging, signal detection, time-delay estimation, white-light scanning interferometry, etc., can be studied in [222–224]. For the speech applications, CTEO is exploited for far-field ASR and signal detection (because the TEO requires a single channel input, whereas CTEO exploits the spatial diversity information from multiple channels) [186, 225]. Furthermore, modulation features extracted using CTEO are also explored for ASR [187, 226]. The selected chronological progress for the proposed CTECC_{max} feature set is shown in Figure 4.13.

While there are other potential approaches, such as beamforming methods, to exploit the spatial diversity of speech waves via microphone array. In particular, state-of-the-art Minimum Variance Distortionless Response (MVDR) is an efficient beamformer that maximizes the array gain, which is a measure of the increase in

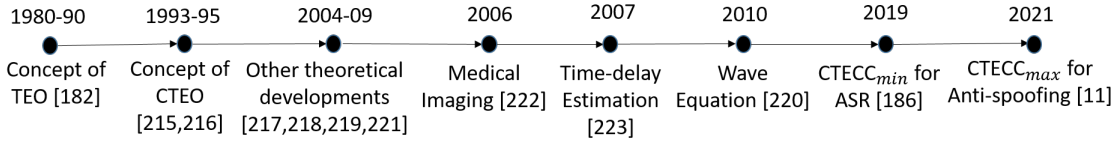


Figure 4.13: Selected Chronological Progress to Develop CTECC_{max} Feature Set for Anti-Spoofing.

signal-to-noise ratio (SNR) that is obtained from the microphone array rather than a single microphone channel. However, MVDR suffers from a serious limitation of having low directivity factor in low frequency regions that carry important discriminating acoustic cues for genuine *vs.* replay spoof. In addition, the filtering process in MVDR distorts the characteristics of speech spectrum [186]. To that effect, we employ CTEO framework to exploit spatial diversity in array. In particular, this study proposes the CTECC_{max} for replay SSD on VAs that process multi-channel inputs. The earlier study proposed CTECC_{max} feature set for far-field ASR, where it considered the minimum cross-Teager energies between the sub-band filtered signals to capture the noise-robust feature representation. However, in our work, we select the most noisy transmission channel in order to track maximum distortions due to replay conditions, i.e., it provides the more discriminative information of the underlying acoustical environment in which replay attack is mounted on VAs. This is the key novelty of the proposed CTECC_{max} feature set for replay SSD task. The proposed approach captures the maximum distortion and hence, it is acronymed as CTECC_{max}. Since, the earlier version of CTCCC was utilized for far-field ASR to capture the minimum distortion, we abbreviate it as CTECC_{min}. The experiments were performed on the dataset configurations as suggested in a recent study [12], where complex deep learning architecture has been utilized for extracting the information from multi-channel speech signal to build the SSD system. It has been observed that the proposed CTECC_{max} feature set outperforms the other existing feature sets, as well as the complex deep learning architecture proposed in [12]. Further theoretical analysis, experimental setup, and results are discussed in subsequent Sections.

4.5.1 Cross-Teager Energy Operator (CTEO)

As discussed in Section 4.3, TEO was originally developed for single channel signals [182]. Hence, to track the cross-Teager energies between two channels, CTEO is developed in [215], and can be denoted as $\Psi_{cr}\{\cdot\}$. CTEO is a non-linear quadratic operator, which estimates the relative rate of change of energies be-

tween signals. The Cross-Teager Energy (CTE) between the two *real-valued* signals, $x(t)$ and $y(t)$ in continuous-time domain is represented as [186]:

$$\Psi_{cr}\{x(t), y(t)\} = \dot{x}(t)\dot{y}(t) - x(t)\ddot{y}(t), \quad (4.45)$$

$$\Psi_{cr}\{y(t), x(t)\} = \dot{y}(t)\dot{x}(t) - y(t)\ddot{x}(t). \quad (4.46)$$

The concept of TEO and CTEO can also be derived using *Lie bracket* [216]. Instantaneous differences in the relative rate of change between two signals $x(t)$ and $y(t)$ can be measured via their Lie bracket ($[\cdot]$) as follows [216]:

$$[x(t), y(t)] = \dot{x}(t)y(t) - x(t)\dot{y}(t). \quad (4.47)$$

In eq. (4.47), if $y(t) = \dot{x}(t)$ then it becomes TEO as given in eq. (4.4). If $x(t)$ and $y(t)$ represent displacements in some generalized motions, then the quantity $[x(t), \dot{y}(t)] = \dot{x}(t)\dot{y}(t) - x(t)\ddot{y}(t)$ has dimensions of energy and hence, it can be referred to as *cross-Teager energy* between $x(t)$ and $y(t)$. Hence, eq. (4.48) is supposed to estimate the cross-Teager energy between two signals, $x(t)$ and $y(t)$ and can be referred to as CTEO [186]:

$$\Psi_{cr}\{x(t), y(t)\} = \dot{x}(t)\dot{y}(t) - x(t)\ddot{y}(t). \quad (4.48)$$

From eq. (4.46), we have $\Psi_{cr}\{x(t), x(t)\} = \Psi\{x(t)\}$, $\Psi_{cr}\{y(t), y(t)\} = \Psi\{y(t)\}$ and $\Psi_{cr}\{0, x(t)\} = \Psi_{cr}\{x(t), 0\} = 0$, $\Psi\{b\} = 0$, where b is a constant. From eq. (4.45) and eq. (4.46), the non-commutative property of CTEO is observed, i.e., $\Psi_{cr}[x(t), y(t)] \neq \Psi_{cr}[y(t), x(t)]$ [215], [225]. Using eq. (4.45), the *average CTEO* ($\Psi_{cr}^{avg}[\cdot]$) between the continuous-time *real-valued* signals is estimated as [225]:

$$\Psi_{cr}^{avg}\{x(t), y(t)\} = \frac{1}{2}[\Psi_{cr}\{x(t), y(t)\} + \Psi_{cr}\{y(t), x(t)\}]. \quad (4.49)$$

However, the definition of CTEO can be extended to complex-valued signals as given in [218]. Furthermore, for the discrete-time signals $x(n)$ and $y(n)$, average cross-Teager energies are estimated as [173]:

$$\begin{aligned} \Psi_{cr}^{avg}\{x(n), y(n)\} = & x(n)y(n) - 0.5[x(n+1)y(n-1) \\ & + x(n-1)y(n+1)]. \end{aligned} \quad (4.50)$$

From eq. (4.50), the excellent time resolution of the CTEO can be observed. Subsequently, the later part of the paper deals with the real-valued continuous-time do-

main representation of speech signal, which can be further extended to discrete-time domain.

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

$$x_i(t) = s_i(t) + n_i(t), \quad i = 1, 2, \dots, N. \quad (4.51)$$

It should be noted that $s_i(t)$, i.e., the speech recording at the i^{th} microphone is indeed dependent on geometry of microphones (details mentioned in Table 3.13) and it should be noted that reverberation phenomenon is not considered via eq. (4.51) [186, 187]. For ASR task, it makes sense to choose the microphone-pair with the lowest cross-Teager energies as it will minimize the environmental distortion (including reverberation) and thus, increases performance of ASR system. On the other hand, this environmental distortion, in particular, reverberation could serve as an important acoustic cue to discriminative between genuine and replay attack should we choose relatively maximum energy in microphone-pair and thus, increases replay SSD performance.

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

$$x_{i_j}(t) = x_i(t) * g_j(t), \quad j = 1, 2, \dots, L, \quad (4.52)$$

where $*$ represents the convolution operation and $x_{i_j}(t)$ represents the subband filtered signal obtained for the i^{th} channel and j^{th} subband filter (having $g_j(t)$ as impulse response) in the filterbank. Considering two sensor inputs (p, q) and j^{th} subband filter of the filterbank, the CTEO will be expressed as:

$$\Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\} = \dot{x}_{p_j}(t)\dot{x}_{q_j}(t) - x_{p_j}(t)\dot{x}_{p_j}(t). \quad (4.53)$$

From the eq. (4.11), eq. (4.51), and eq. (4.53), we obtain:

$$\begin{aligned} \Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\} &= \Psi\{s_j(t)\} + \Psi_{cr}\{n_{p_j}(t), n_{q_j}(t)\} \\ &+ \Psi_{cr}\{s_j(t), n_{q_j}(t)\} + \Psi_{cr}\{n_{p_j}(t), s_j(t)\}. \end{aligned} \quad (4.54)$$

The replay noise is represented by the last three terms on the Right-Hand Side (RHS) of eq. (4.54). Taking expectation operator ($E[\cdot]$) on eq. (4.54), we get:

$$E[\Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\}] = E[\Psi\{s_j(t)\}] + E[\Psi_{cr}\{n_{p_j}(t), n_{q_j}(t)\}]. \quad (4.55)$$

The last two terms of RHS side of eq.(4.54) are zero-mean and hence, the ex-

peptation operator is zero [186]. However, the second term represents the error in eq. (4.55) [31]. Hence, the modified equation is given as:

$$E[\Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\}] = E[\Psi\{s_j(t)\}] + error, \quad (4.56)$$

where $error = E[\Psi_{cr}\{n_{p_j}(t), n_{q_j}(t)\}]$. Let us denote Γ the concentration of noise power within the subband filter's passband. Using Cauchy-Schwartz inequality for two random variables X and Y , we have [227, 228]:

$$|E(XY)|^2 \leq E(X^2)E(Y^2), \quad (4.57)$$

where (XY) is the inner product between the random variables X and Y . Therefore, using eq. (4.57), the relation between the noise power (proof is given in Appendix), we obtain:

$$|\Gamma_{(pq)_j}|^2 \leq \Gamma_{p_j}\Gamma_{q_j}, \quad (4.58)$$

where Γ_{p_j} is the noise power concentration of the j^{th} subband and p^{th} channel. Moreover, Γ_{p_j} is proportional to the $error$ term in eq. (4.56), where the $error$ term is the varying, whereas the source signal through the bandpass filter remains the same throughout the analysis. For ASR, the desirable speech signal representation should contain the least amount of noise component. Hence, the representation with minimum $error$ is chosen for ASR application as explained in [187]. Whereas, for replay SSD, it is necessary to emphasize the distorted channel information and hence, we have chosen the channels, which corresponds to maximum $error$ in eq. (4.56). By maximizing the $error$, the additional acoustical representation can be obtained. With respect to the analysis of CTEO, we have ${}^N C_2$ possibilities of channel-pairs for each i^{th} subband. Estimating average CTE for all the channel-pairs and then choosing the one with the highest average energy is a feasible, however, it is computationally expensive approach. To reduce the computational complexity, the two channels with the highest average Teager energy can be chosen and CTE between those two channels can be utilized for further representation. Furthermore, among the set of the one average CTE and two Teager energies, the subband filtered signal with the maximum energy is selected for classification between genuine and replay utterances, namely, Maximum Energy Signal (MES). Mathematically, MES can be represented as [11]:

$$MES = \max_{(p,q)} (E[\Psi_{cr}^{avg}\{x_{p_j}(t), x_{q_j}(t)\}], E[\Psi\{x_{p_j}(t)\}], E[\Psi\{x_{q_j}(t)\}]). \quad (4.59)$$

From eq. (4.59), the MES contains the maximum distortions, such as acousti-

cal environment and intermediate device responses, these non-linearities are captured by the MES. Hence, for the replay SSD, the MES is selected for further processing.

4.5.2 CTECC_{max} Feature Extraction Procedure

Figure 4.14 shows the functional block diagram of the CTECC_{max} feature set developed for replay SSD task on VAs. The dataset consists of the recordings from a variety of VAs having various sampling rates (as shown in Table 3.13). Hence, the input speech of each channel is re-sampled at 16 kHz. Each of the signal from N -channel microphone array is processed through a Gabor filterbank, which possess excellent time-frequency resolution (because the Fourier transform of a Gaussian function is also a Gaussian. Further, Gaussian belongs to the class of infinitely differentiable functions, in particular, $\{g_j(t) \in \mathbb{C}^\infty\}, j \in [1, 160]$ and hence, they have faster decay in frequency-domain [200]). The Gabor filterbank consist of linearly-spaced 160 subband filters and hence, we obtain 160 subband filtered signals for each channel. Then, a TEO profile for each subband filtered signal is obtained. Furthermore, average of the TEO_{nj} are compared, where $n \in [1, N]$ (N corresponds to number of channels in the microphone array as shown in eq. (4.51)) and $j \in [1, 160]$. Then, two channels p and q are selected such that they have maximum average TEO. Using eq. (4.50) on the p and q , the average CTEO is estimated. Windowing is performed on the subband filtered signal with a window size of 25 ms and window shift of 10 ms, which provides m frames. Averaging on each frame is performed, which provides the average energy for a frame in consideration. Then logarithm operation is performed, which is followed by Discrete Cosine Transform (DCT) to obtain the cepstral representation. Initial 70 DCT (static) features are concatenated with dynamic Δ and $\Delta\Delta$ coefficients, which results in 210-dimensional (D) CTECC_{max} feature vector. It can be clearly observed that for a single channel case, the proposed CTECC_{max} feature set would result in TECC feature set representation as there would be no channel selection procedure to follow after estimating the TEO profile for each subband [106]. The MATLAB pseudocode for implementation of CTECC_{max} is shown in Algorithm 3. The original CTECC_{min} feature set, which was proposed for ASR to minimize the distortions, can be computed by replacing $maxk(\cdot)$ in line 9 and line 13 in Algorithm 3 to $mink(\cdot)$.

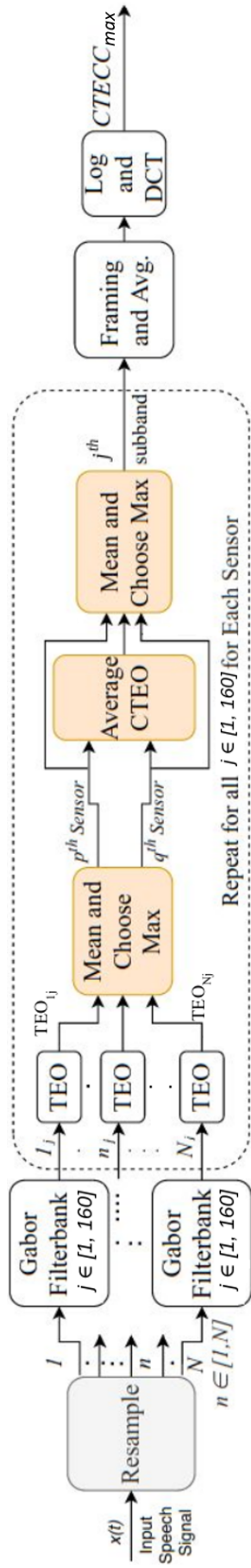


Figure 4.14: Functional Block Diagram of Proposed CTECC_{max} Feature Extraction. After [11, 13].

Algorithm 3 MATLAB Pseudocode of Proposed CTECC_{max} Feature Set Extraction. After [11].

1. $x = \text{ReSample}(x, 16000)$, resampling of signal to 16 kHz,
 2. $f\text{bank}G = \text{Gabor_fbank}(Q, bw)$, construct the Gabor filterbank having Q subband filters with
bandwidth bw ,
 3. **for** $i = 1 : Q$ **do**,
 for $j = 1 : N$ **do**, N represents number of channels in microphone array,
 $y(j, i, :) = \text{filter}(f\text{bank}G)(i, :, 1, x(j, :))$, subband filtering using i^{th} subband filter,
 $T_{\text{subband}}(j, i, :) = \text{TEO}(y(j, i, :))$, estimate energy using TEO,
 end for
 $T_{\text{avg}}(1 : N) = \text{mean}(T_{\text{subband}}(j, i, :))$, estimate average for each subband channel,
 $[\sim, k] = \text{maxk}(T_{\text{avg}}(1 : N), 2)$, sorting *w.r.t.* energy,
 $C_{\text{subband}}(i, :) = \text{cross_TEO}_{\text{avg}}(y(k(1), i, :), y(k(2), i, :))$, estimate average-CTE between two
 highest TE signals
 $C_{\text{avg}}(i) = C_{\text{subband}}(i, :)$
 $\text{feat1}(1 : 3, :) = [T_{\text{subband}}(k(1), i, :); T_{\text{subband}}(k(2), i, :); C_{\text{subband}}(i, :)]$
 $[\sim, k2] = \text{maxk}(T_{\text{avg}}(k(1)), T_{\text{avg}}(k(2)), C_{\text{avg}}(i))$, sorting among two TE and one CTE,
 $\text{feat}(i, :) = \text{feat1}(k2, :)$, choose highest among two TE and one CTE,
 $C_{\text{frames}} = \text{enframe}(\text{feat}(i, :), \text{win_len}, \text{win_shift})$, framing on selected feature vector,
 $C_{\text{avg}}(i, :) = \text{mean}(C_{\text{frames}})$, Averaging over each frame,
 end for
 4. $C_{\log} = \log(\text{abs}(C_{\text{avg}}))$,
 5. $C_{\text{static}} = \text{DCT}(C_{\log})$, static coefficients,
 6. $C_{\Delta} = \text{delta}(C_{\text{static}})$, velocity coefficients,
 7. $C_{\Delta\Delta} = \text{delta}(C_{\Delta})$, acceleration coefficients,
 8. $\text{CTECC} = [C_{\text{static}}; C_{\Delta}; C_{\Delta\Delta}]$, CTECC_{max} feature set.
-

4.5.3 Experimental Setup

The experiments for the proposed CTECC_{max} feature set were performed using ReMASC dataset with configuration shown in Table 3.15. The experiments are also extended for the recordings from the individual devices. For this case, the dataset for each of the device is partitioned with 90 % training and 10 % Dev set with overlapping speakers. The device-specific partition of the dataset is given in Table 3.13. CQCC (Baseline), MFCC, LFCC, and CTECC_{max} feature sets are utilized in this work along with GMM, CNN, and LCNN as classifiers. Furthermore, the various systems are evaluated using the % EER as an evaluation metric.

4.5.4 Spectrographic Analysis

Panel 1 and Panel 2 of Figure 4.15 shows genuine and replay speech signal corresponding to the same utterance "Hey Cortana, remind me to pick up Chick-fil-A", respectively. Figure 4.15(a) and Figure 4.15(b) shows the time-domain signal, and its corresponding STFT spectrogram, respectively. However, Figure 4.15(c) and

Figure 4.15(d) shows the spectrogram representations for TECC and $CTECC_{max}$ feature sets, respectively. Spectrogram representations for TECC and $CTECC_{max}$ feature sets refers to the feature representation *via* ESD obtained just before applying the DCT operation. It can be observed from Figure 4.15(b), Figure 4.15(c), and Figure 4.15(d) that the STFT-based representations have the less difference in the speech *vs.* silence regions. Hence, it is not able to effectively emphasize the speech information, however, it is more effectively emphasized by the ESD corresponding to $CTECC_{max}$ feature set. This might be the reason that the proposed $CTECC_{max}$ feature set performs relatively better for replay SSD task, than the other feature sets used in this study.

$CTECC_{max}$ has two advantages, namely, first it is capturing the characteristics of utterance in more clear way. This can be observed with the absence of background noise in the Figure 4.15(d) (for $CTECC_{max}$) as compared to corresponding regions of ESD in Figure 4.15(c). However, in Figure 4.15(d) (Panel II) inspite of background noise supression capability of the proposed $CTECC_{max}$ feature set, we observe a distinct and continuous band of energy in very low frequency region across the *entire* time duration. This means that the continuous band of energy is only due to the replay effect, making it a distinguishing acoustic cue for replay SSD task. Notably, this band of ESD is the most distinct in the $CTECC_{max}$ -based ESD representation as shown in Figure 4.15(d) (Panel II) as compared to their STFT and TECC representation in Figure 4.15(b) and Figure 4.15(c), respectively.

Second, noise suppression capability of the $CTECC_{max}$ is observed from the spectrogram. It can be observed that there is a sudden discontinuity of background noise around 6000 Hz, in particular, for STFT and TECC-based ESD representation as shown in Figure 4.15(b) and Figure 4.15(c), respectively. However, the background noise as well as sudden discontinuities in noise are suppressed significantly in the corresponding $CTECC_{max}$ representation.

4.5.5 Results on Individual Systems and Their Fusions

Results for CQCC (Baseline), MFCC, LFCC, and $CTECC_{max}$ feature sets are shown in Table 4.17 for the dataset configuration in Table 3.15. For each of these feature sets, % EER is shown on Dev and Eval dataset using GMM as the back-end Bayesian classifier. It can be observed that we obtained absolute reduction in EER of 4.11 % and 7.38 % on Dev and Eval sets, respectively, as compared with the baseline CQCC-GMM system. It can also be observed that $CTECC_{max}$ performs better than $CTECC_{min}$, which validates our approach of maximizing the distortions present in the signal in order to identify degradation in speech utter-

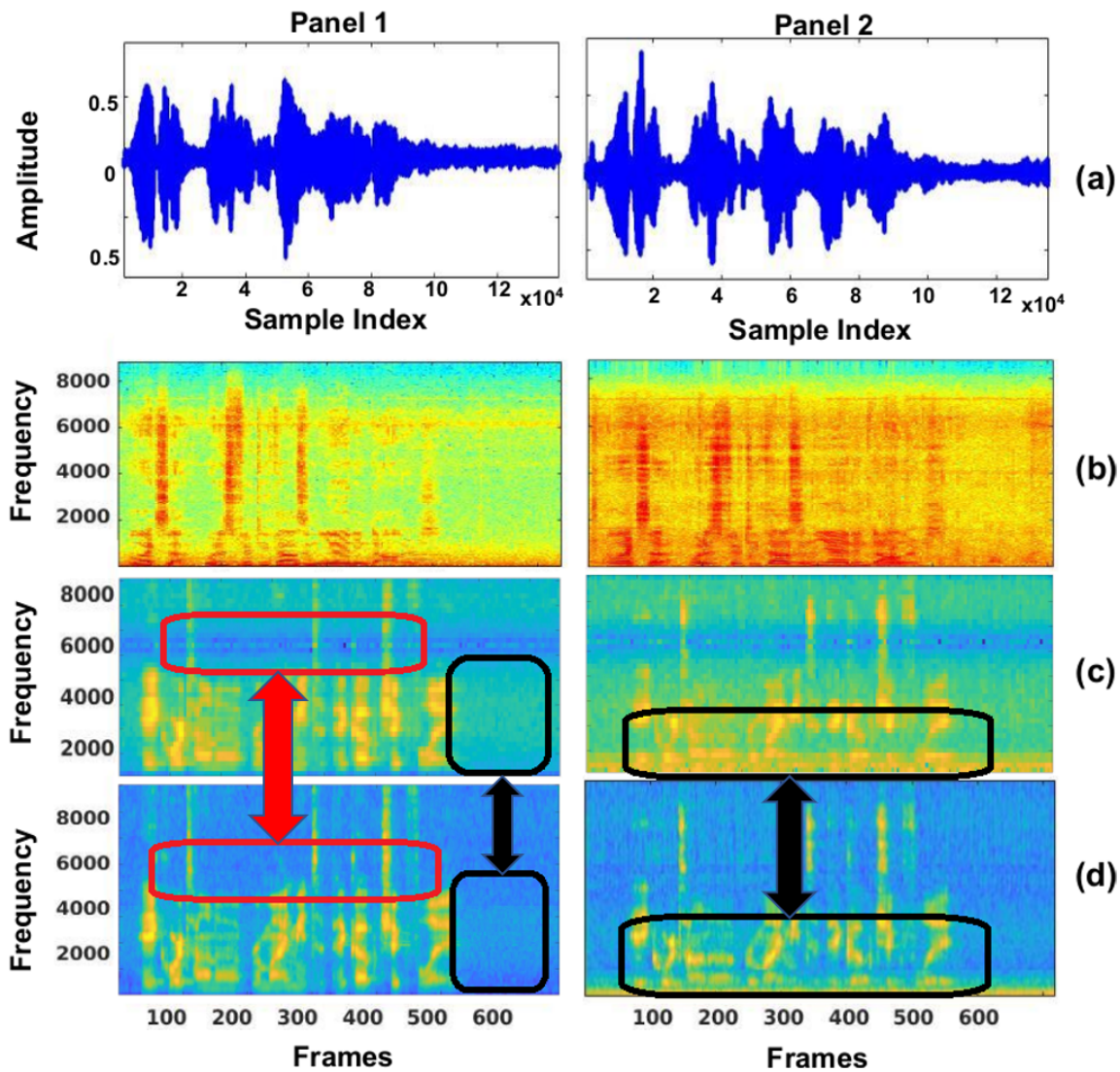


Figure 4.15: Spectrogram Plot of $CTECC_{max}$ (Panel I) *vs.* CQCC (Panel II) Feature Sets : (a), (b) for Genuine Speech Signal, and (c), (d) for Spoofed Speech Signal. After [11].

ance due to replay configurations. Experiments are also performed using LCNN with CQCC and $CTECC_{max}$ feature set, showing the similar trends in results as in GMM. Furthermore, EER is reduced to 13.49 % and 13.99 % using score-level fusion of $CTECC_{max}$ -GMM and $CTECC_{max}$ -LCNN systems on Dev and Eval sets, respectively, indicating both of these systems capture complementary information.

Table 4.17: Results (in % EER) on ReMASC Dataset. After [11].

Feature Set	Dev	Eval
CQCC-GMM	20.57	23.31
CQCC _{max} -GMM	18.65	22.67
LFCC-GMM	23.44	21.79
LFCC _{max} -GMM	21.73	22.12
MFCC-GMM	36.43	31.53
CTECC _{min} -GMM	20.56	17.30
CTECC _{max} -GMM (A)	16.46	15.93
CQCC-LCNN	22.31	25.88
CTECC _{max} -LCNN (B)	16.87	19.70
A + B	13.49	13.99

'+' denotes score-level fusion.

4.5.6 Results Obtained on Environment-Dependent and Environment-Independent Scenarios

For environment-dependent scenario, the target environment is already seen by the defense model. In this case, each environment is partitioned into two disjoint and speaker-independent sets of roughly the same size. The results obtained using CQCC and CTECC_{max} feature sets with the application of CMVN, are reported in Table 4.18. It is noteworthy that in this scenario, CMVN had improved the performance of both the systems. It can be observed from Table 4.18 that in case of Env-A and Env-B, CTECC_{max} shows absolute reduction of 10.23 % and 15.84 %, respectively, over CQCC feature set possibly because replay configurations in Env-A and Env-B are more noise-dominant in the sense that both genuine and spoof utterances have greater noise imbalances. Hence, these environments show better discrimination. This observation is well supported by experimental results obtained on Env-C and Env-D, where % absolute reduction obtained is just 3.05 % and 5.8 %, respectively, over CQCC feature set, possibly due to the fact that for these replay configurations, background noise in genuine and spoof configuration is more or less the same.

Environment-independent experiments, on the other hand, are performed by training defense models on any of the three environments, and tested on the fourth unseen environment. Results suggest that the performance of CTECC_{max} feature set is better than the baseline system on environment A, B, and C, which shows that the proposed cross-Teager energy-based features provides stronger feature discrimination ability to the unseen environment conditions. However,

Table 4.18: Results in % EER for Environment-Dependent *vs.* Environment-Independent Case on ReMASC Dataset. After [11].

	Feature Set	Env-A	Env-B	Env-C	Env-D
Environment Dependent	CQCC	23.27	42.62	12.96	15.85
	CTECC _{max}	13.04	26.78	9.91	10.05
Environment Independent	CQCC	35.65	40.89	35.95	49.99
	CTECC _{max}	28.49	34.68	32.52	49.99

both baseline CQCC, and CTECC_{max} feature set perform poorly on environment D, which indicates that it is completely different from A, B, and C. Hence, this aspect needs further investigation and is an open research issue.

4.5.7 Results using Individual Recording Devices

For performance analysis on individual devices, we use the core set for training and Dev set, whereas Eval set is utilized for testing. The training and Dev sets consist of 90 % and 10 % of the core set, respectively, of each recording devices with overlapping speakers. Additionally, for the data collection, various recording devices are utilized with different specification as shown in Table 3.13 [9]. The performance of the proposed CTECC_{max} feature set is compared against the state-of-the-art features, such as MFCC, CQCC, LFCC, and TECC, extracted from the first channel of the microphone array. The GMM, CNN, and LCNN classifiers are utilized for performance analysis on individual devices. Initially, the parameters of the proposed CTECC_{max}, namely, number of subband filters in the filterbank and number of dimensions of the feature vector are fine-tuned for optimal performance using several experiments. It can be observed from Figure 4.16 that the relatively better results are obtained for all the four devices (D1-D4) with 160 number of subband filters. Furthermore, with 160 number of subband filters in the filterbank, experiments are extended by varying the dimension of the feature vector, which includes static, Δ , and $\Delta\Delta$ coefficients. It can be observed from Figure 4.17 that 210-D feature set produce the optimum % EER for all the four devices. Moreover, Figure 4.18 represents the performance of the 210-D feature set *w.r.t.* number of mixtures used in GMM classifier. It can be observed from Figure 4.18 that the optimum performance is obtained using 512 mixtures in GMM. Further experiments are performed with the CTECC_{max} feature set extracted using 160 number of subband filters and 210-D feature representation, which includes static, Δ , and $\Delta\Delta$ features.

The comparison of the proposed CTECC_{max}-GMM architecture and the deep

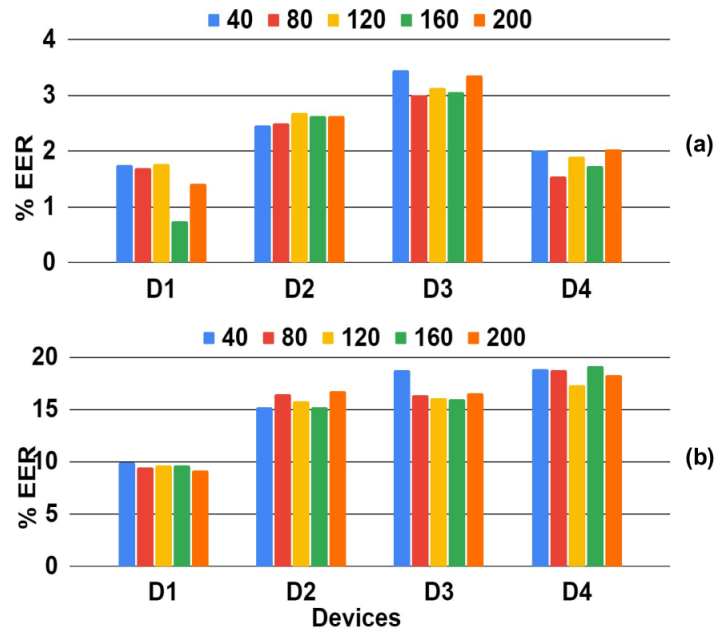


Figure 4.16: Results using $CTECC_{max}$ w.r.t Number of Subband Filters used in Gabor Filterbank: (a) Dev Set, and (b) the Eval Set.

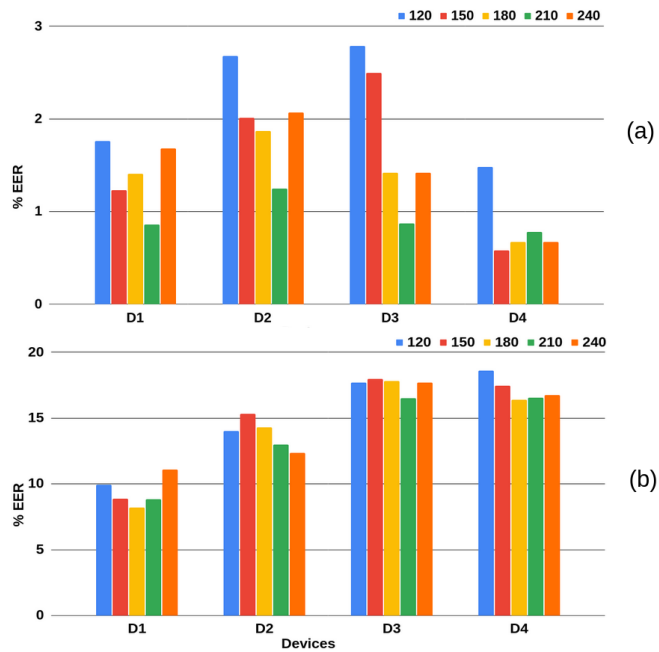


Figure 4.17: Results w.r.t Dimension of $CTECC_{max}$ Feature Vector: (a) Dev Set, and (b) Eval Set.

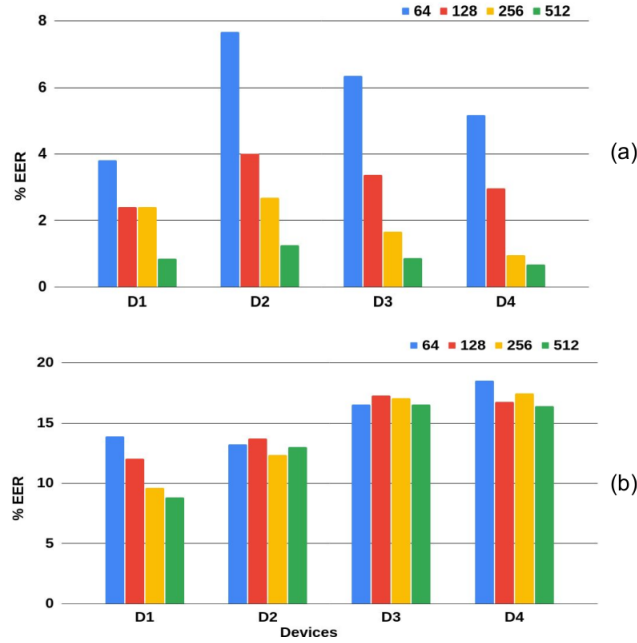


Figure 4.18: Results *w.r.t* Number of Mixtures in GMM using $CTECC_{max}$: (a) Dev Set, and (b) Eval Set.

learning-based approach utilized in [12] is shown in Table 4.19 *w.r.t.* the number of channels utilized in each device. Both the architectures exploit the multi-channel information representation for replay SSD task. The architecture in [12] was considered as a baseline architecture in our work. As shown in Table 3.13, the devices D1, D2, D3, and D4 consists of 2, 4, 6, and 7 channels, respectively. It can be observed from the Table 4.19 that the proposed $CTECC_{max}$ feature set performs relatively better than the baseline architecture for devices D1, D2, and D4, when all the available channels in the microphone array were utilized for feature representation. Whereas, the comparable performance is observed for the device D3.

Furthermore, the performance comparison of the proposed $CTECC_{max}$ feature set with the other features using GMM classifier is shown in Table 4.20. It can be observed that the proposed $CTECC_{max}$ feature set performs better than the other feature sets, except for device D1. In addition, similar trends in results are observed for all the feature sets using deep learning-based architectures, such as CNN and LCNN as shown in Table 4.21 and Table 4.22, respectively. On the whole, the proposed $CTECC_{max}$ feature set is useful representation for the replay SSD for VAs, where multi-channel information can be exploited. Furthermore, the LLR score distribution of genuine *vs.* spoof speech utterances on Eval set for device D4 is shown in Figure 4.19. The common area under the two Gaussian-like curves in Figure 4.19 is the probability of misclassification [212]. We call this

area as *intersecting area*. The lesser the intersecting area is, the lesser the probability of misclassification is and hence, the lesser the % EER for such a feature set will be. It can be observed from Figure 4.19 that the proposed CTECC_{max} feature set occupies relatively the least amount of area, indicating its relatively better discrimination power than the other feature sets.

Table 4.19: Results (in % EER) for Comparison of CTECC_{max} with Existing Architecture in [12] on Eval set for Various Devices. After [13].

Device	Channels Utilized for Replay SSD							
	1		2		3		4	
	[12]	CTECC _{max}	[12]	CTECC _{max}	[12]	CTECC _{max}	[12]	CTECC _{max}
D1	16.6	22.0	14.9	8.84	-	-	-	-
D2	23.7	25.80	19.5	15.74	16.7	16.33	15.4	13.01
D3	23.7	24.63	19.1	17.31	17.6	19.755	17.0	19.10
D4	27.5	29.79	21.5	21.47	20.6	21.19	21.3	20.3

Device	Channels Utilized for Replay SSD							
	5		6		7		-	-
	[12]	CTECC _{max}	[12]	CTECC _{max}	[12]	CTECC _{max}	-	-
D1	-	-	-	-	-	-	-	-
D2	-	-	-	-	-	-	-	-
D3	17.1	19.71	16.5	16.53	-	-	-	-
D4	20.7	20.25	19.9	21.15	19.8	16.41	-	-

Table 4.20: Results (in % EER) on Dev and Eval Set *w.r.t.* Various Feature Sets and Devices using GMM Classifier. After [13].

Device	D1		D2		D3		D4	
	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
MFCC	9.16	7.98	16.87	26.02	19.71	20.37	12.23	26.99
CQCC	2.88	11.9	4.59	28.68	4.07	23.91	1.88	29.392
LFCC	2.26	8.04	3.45	20.09	4.75	19.32	3.43	23.18
TECC	9.15	22.0	12.34	25.80	10.85	24.63	13.72	29.79
CTECC _{min}	0.429	8.00	1.90	20.07	2.07	19.19	1.04	19.68
CTECC _{max}	0.86	8.84	1.25	13.01	0.87	16.53	0.67	16.41

4.5.8 Detection Error Trade-off (DET) Curves

The performance of proposed feature set is also evaluated using the DET curves for various feature sets because it gives SSD performance at various operating points of the SSD system. Figure 4.20 shows the DET curves obtained for the devices D1, D2, D3, and D4 for Dev and Eval sets. Figure 4.20(a) and Figure 4.20(e) shows the DET curves for device D1 on Dev and Eval set, respectively. Similarly, (Figure 4.20(b), Figure 4.20(f)), (Figure 4.20(c), Figure 4.20(g)), and (Figure 4.20(d),

Table 4.21: Results (in % EER) on Dev and Eval Set *w.r.t.* Various Feature Sets and Devices using CNN Classifier. After [13].

Device	D1		D2		D3		D4	
Feature Set	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
MFCC	3.93	14.89	5.91	24.80	8.63	21.65	7.18	24.86
CQCC	6.29	17.92	9.14	36.06	16.50	32.30	13.70	40.90
LFCC	4.72	15.40	6.89	36.70	5.75	26.50	7.18	25.94
TECC	19.48	29.4	19.30	27.21	15.38	26.05	26.94	29.90
CTECC _{min}	2.36	14.39	5.37	27.54	2.87	27.81	1.79	22.82
CTECC _{max}	2.76	11.11	3.76	23.37	2.87	25.17	3.59	22.82

Table 4.22: Results (in % EER) on Dev and Eval Set *w.r.t.* Various Feature Sets and Devices using LCNN Classifier. After [13].

Device	D1		D2		D3		D4	
Feature Set	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
MFCC	7.08	12.87	11.82	29.38	16.54	27.46	16.16	34.41
CQCC	10.23	18.43	16.67	35.39	17.26	25.70	13.77	37.47
LFCC	12.59	20.20	16.66	38.56	17.98	28.87	18.56	39.07
TECC	29.91	32.32	30.64	35.05	30.21	33.45	36.51	38.91
CTECC _{min}	9.44	16.41	4.83	25.70	8.63	27.4	4.79	30.45
CTECC _{max}	3.56	16.91	4.32	28.21	9.35	21.3	6.58	24.34

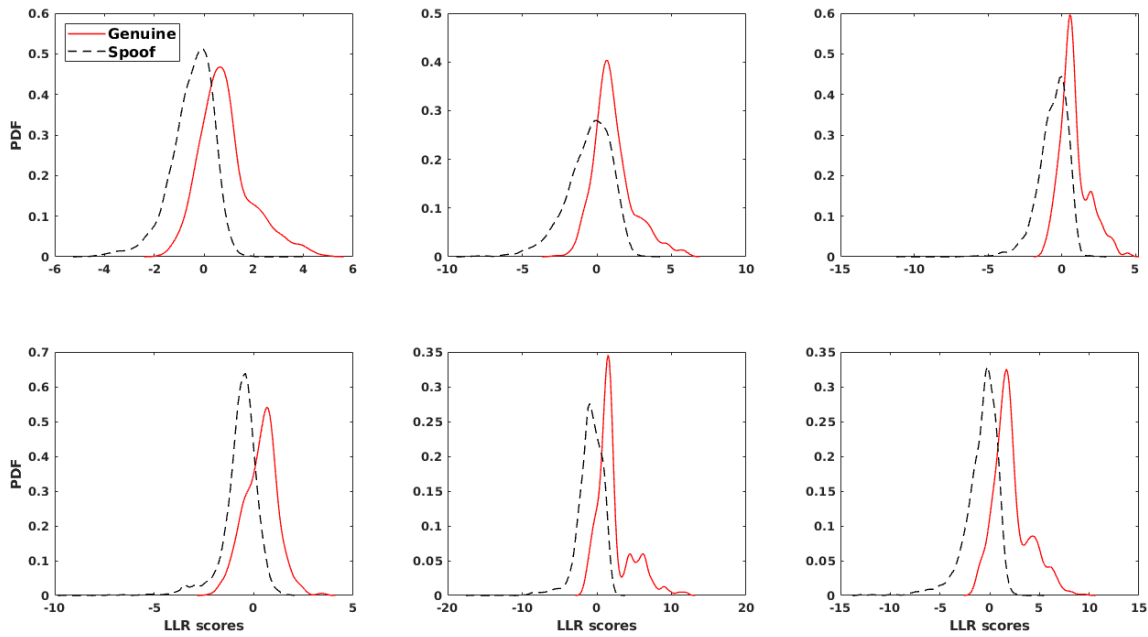


Figure 4.19: LLR Scores Distribution on Eval Set of D4: (a) MFCC, (b) CQCC, (c) LFCC, (d) TECC, (e) CTECC_{min}, and (f) CTECC_{max}. After [13].

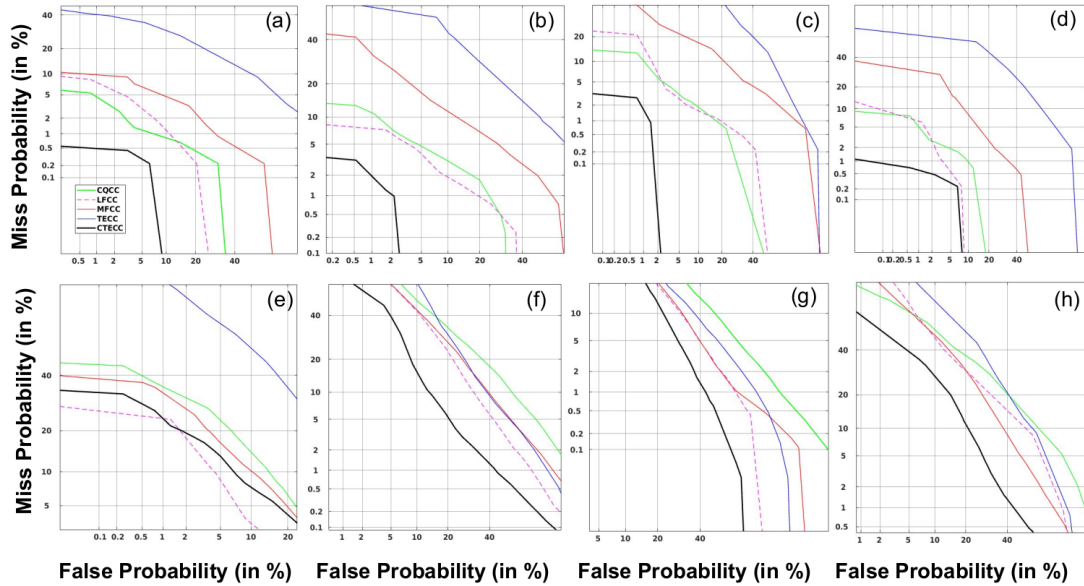


Figure 4.20: DET Curves Obtained for Various Feature Sets as Shown in Legends. Figure 4.20(a) and Figure 4.20(e) shows the DET Curves for Device D1 on Dev and Eval Set, Respectively. Similarly, (Figure 4.20(b), Figure 4.20(f)), (Figure 4.20(c), Figure 4.20(g)), and (Figure 4.20(d), Figure 4.20(h)) shows the DET Plots for Device D2, D3, and D4 on (Dev, Eval) Set, Respectively. The Legend shown in Figure 4.20(a) is Similar for Remaining DET Plots.

Figure 4.20(h) shows the DET plots for device D2, D3, and D4 on (Dev, Eval) set, respectively. It can be observed that the $CTECC_{max}$ outperforms the other feature sets for all the devices (except device D1) for MFCC, and LFCC feature set at all the operating points of SSD system.

The performance of the proposed $CTECC_{max}$ feature set is summarized in summary (i.e., Section 4.7). In the next Section, the development and the performance of the proposed CFCCIF-ESA feature set is discussed.

4.6 CFCCIF-ESA Feature Set

4.6.1 Proposed CFCCIF-ESA Feature Set

As shown in the Figure 4.21, the proposed CFCCIF-ESA feature set is an adaptation of the CFCC feature set, which extracts the magnitude information from the subband filtering using cochlear filterbank. The CFCC feature set was developed for speaker recognition task, and shows the relatively better performance than MFCC, Perceptual Linear Prediction (PLP), and relative spectral-PLP (RASTA-PLP) features under noisy or signal degradation conditions [33]. Furthermore, the IF information is incorporated along with magnitude information, to adapt

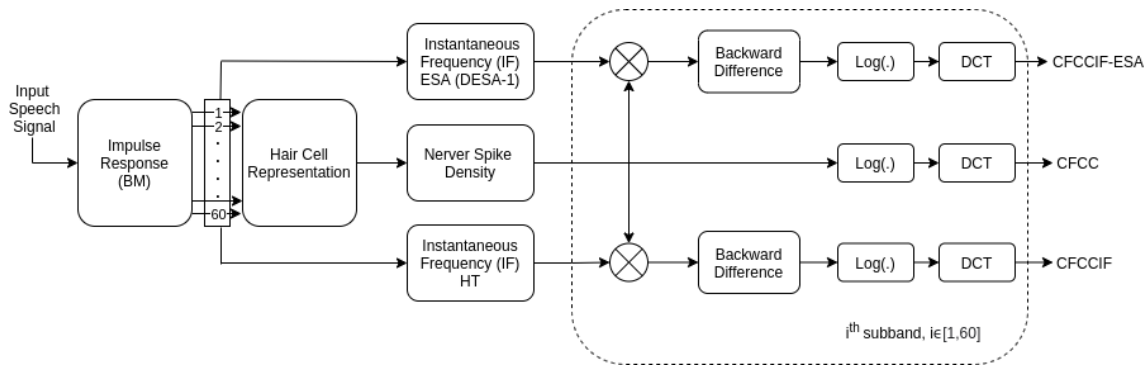


Figure 4.21: Functional Block Diagram of the CFCC, CFCCIF, and Proposed CFCCIF-ESA Feature Set. After [1, 33].

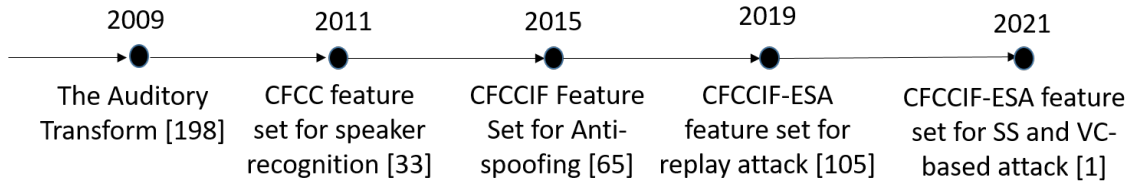


Figure 4.22: Selected Chronological Progress of the Proposed CFCCIF-ESA Feature Set. After [1].

CFCCIF feature set, where IFs were estimated using Hilbert transform approach. Further, IFs are estimated with ESA and embedded with magnitude information, to give CFCCIF-ESA feature set. The CFCCIF and the proposed CFCCIF-ESA feature sets are developed for anti-spoofing. The selected chronological progress for the proposed CFCCIF-ESA feature set is shown in Figure 4.22.

The CFCC feature extraction emulates the human peripheral hearing system and involves cochlear filter-based on auditory transform (AT), hair cell function, non-linearity, and DCT [33]. The AT basically models the traveling wave in the cochlea (in particular, BM), where the decomposition of the sound wave takes place into a set of subband signals [198]. The travelling wave can be modeled by the impulse response function, $\gamma(t) \in L^2(R)$ (i.e., Hilbert space of finite energy signals), which satisfy the following conditions:

- It should be the zero average function, i.e.,

$$\int_{-\infty}^{+\infty} \gamma(t) dt = 0 \Rightarrow \Gamma(\omega)|_{\omega=0} = \int_{-\infty}^{+\infty} \gamma(t) e^{-j\omega t} dt|_{\omega=0} = 0, \quad (4.60)$$

where $\Gamma(\omega)$ is Fourier transform of the $\gamma(t)$, i.e., $\Gamma(\omega) = \mathcal{F}\{\gamma(t)\}$, where $\mathcal{F}\{\cdot\}$ represents the continuous-time Fourier transform (CTFT) operation.

- It suggests that $\Gamma(\omega)$ is bandpass in nature. The bandpass nature of the filter

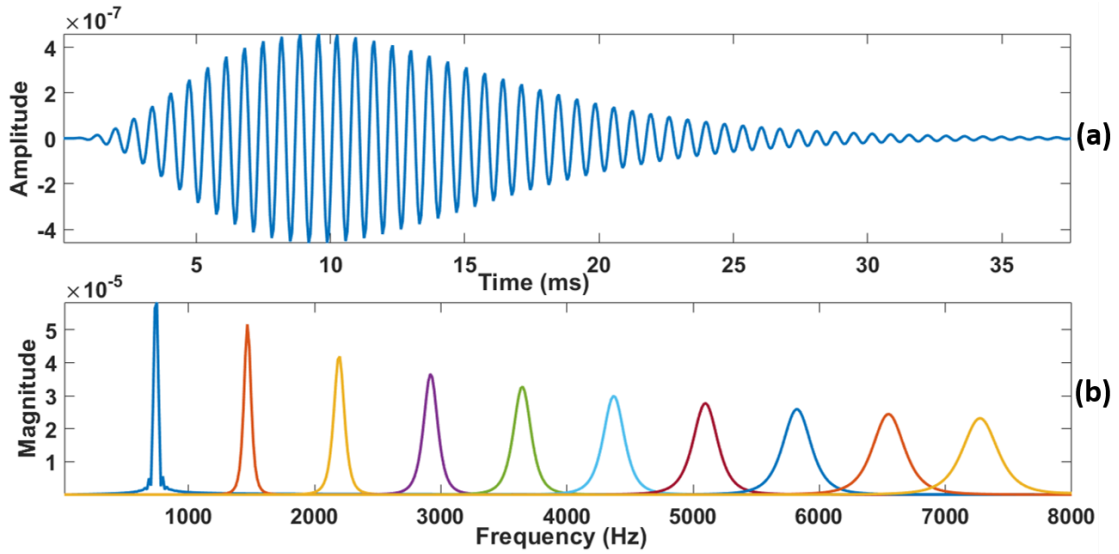


Figure 4.23: (a) Impulse Response of 2nd Subband (Cochlear) Filter, and (b) Corresponding Frequency Response of Cochlear Filterbank Consisting of Ten Subband Filters. After [1].

also helps to approximate the numerical computation [198],

- It decreases to zero on both the sides of t , and it has unit norm, i.e.,

$$\int_{-\infty}^{+\infty} (1 + |t|) |\gamma(t)| dt < +\infty, \text{ and } \|\gamma(t)\| = 1. \quad (4.61)$$

Similar nature is observed in psycho-acoustic experiments with the BM [201]. The impulse response of 2nd subband filter in the filterbank is shown in Figure 4.23.

- Eq. (4.60) indicates that the function $\gamma(t)$ has zero d.c. (average) value, i.e., it is *wavy* in nature, whereas eq. (4.61) indicates that $\gamma(t)$ must decay after a certain short interval of time.

The subband filters in the cochlear filterbank satisfies the above conditions [198]. The AT ($X(u, v)$) of the speech signal, $x(t) \in L^2(\mathbb{R})$ w.r.t the impulse response of the BM ($\gamma(t)$) is given by [33, 198]:

$$X(u, v) = \int_{-\infty}^{+\infty} x(t - \tau) \cdot \gamma_{u,v}^*(\tau) d\tau, \quad (4.62)$$

and

$$\gamma_{u,v}(t) = \frac{1}{\sqrt{u}} \gamma\left(\frac{t-v}{u}\right), \quad (4.63)$$

where the parameter $u \in \mathbb{R}^+$, and $v \in \mathbb{R}$, represents the scaling and translation

parameters, respectively, and * indicates the complex conjugate operation. The scaling parameter u allows to set the center frequency of the subband filters in the cochlear filterbank. The value of u is calculated by using the lowest frequency f_l , and center frequency, f_c , for each subband filter of the cochlear filterbank [198], i.e.,

$$u = \frac{f_l}{f_c}. \quad (4.64)$$

Frequency scale of the filterbank determines the kind of spacing between the center frequencies of the subband filters in the filterbank. The frequency scale of the filterbank can be chosen either ERB, Mel or linear depending upon the application. In this work, we used linear-scale cochlear filterbank by keeping the constant bandwidth for all the subband filters, i.e., it has the *constant* resolution across the entire frequency range. Furthermore, the cochlear filter is defined as in [33]:

$$\gamma_{u,v}(t) = \frac{1}{\sqrt{u}} \left(\frac{t-v}{u} \right)^\alpha \cdot \exp \left[-2\pi f_l \beta \left(\frac{t-v}{u} \right) \right] \times \cos \left[2\pi f_l \left(\frac{t-v}{u} \right) + \theta \right] U(t-v), \quad (4.65)$$

where $U(\cdot)$ represents the unit-step function. The parameters α and β are tuned to set the width and shape, respectively, of the frequency response of the cochlear filter.

Next, we find the frequency response of cochlear filter, $\gamma_{u,v}(t)$. For simplicity and clarity in derivation, let $u = 1$, and $v = 0$ in eq. (4.62) to produce the sample cochlear filter, i.e.,

$$\gamma_{1,0}(t) = \left[t^\alpha \exp(-2\pi f_l \beta t) \right] \cdot \cos(2\pi f_l t + \theta) U(t). \quad (4.66)$$

Applying CTFT on both the sides of eq. (4.66), we get [1],

$$\mathcal{F} \left[\gamma_{1,0}(t) \right] = \frac{1}{2\pi} \mathcal{F} \left[t^\alpha \exp(-2\pi f_l \beta t) u(t) \right] * \mathcal{F} \left[\cos(\omega_c t + \theta) \right], \quad (4.67)$$

where * represents the convolution operation in the frequency-domain (due to modulation theorem for CTFT [42,200]). Rewriting eq. (4.67) as,

$$\Gamma_{1,0}(\omega) = \frac{1}{2\pi} \left(\frac{\alpha!}{(j\omega - 2\pi f_l \beta)^{\alpha+1}} \right) * \left(\pi e^{j\theta} \delta(\omega - \omega_c) + \pi e^{-j\theta} \delta(\omega + \omega_c) \right), \quad (4.68)$$

where $\delta(\cdot)$ represents Dirac-delta function in the frequency-domain. The eq. (4.68) consist of two terms combined by the convolution operation. In particular, the first term represents the shape and size of the cochlear filter in the filterbank,

whereas the second term indicates location of impulses at the center frequency of the cochlear filter. Thus, the spectrum of the subband filter is replicated around the center frequency, ω_c . Furthermore, by varying the value of α and β , the shape and width of the frequency response of the subband filters in the cochlear filterbank can be varied in order to achieve adaptive time-frequency resolution [200].

The subband filtering imitates the bandpass characteristics of the impulse response of the BM of various locations *w.r.t.* place theory of hearing. The inner hair cell is responsible for the automatic movement of BM for neural activities. Various regions in BM move up and down (vibrates) according to the frequency content in the signal. These vibrations further results in the movement of uppermost hair cells, which commences of the neural activity. This neural activity is generated in a single direction by the inner hair cells and hence, it can be effectively modelled by a square function as [33]:

$$H(u, v) = (X(u, v))^2, \quad \forall X(u, v), \quad (4.69)$$

where $X(u, v)$ is the cochlear filterbank output, and $H(u, v)$ is the hair cell function. The output of the hair cell is transformed to electrical signals, which are then sent to the brain by the auditory nerves. The NSD can be used to model the intensity of this electrical signal, which is computed as [33]:

$$NSD(i, n) = \frac{1}{w} \sum_{c=m}^{m+w-1} H(i, c), \quad m = 1, h, 2h, \dots; \forall i, n, \quad (4.70)$$

where w is the window length, n is the frame count, and h is the hop size of the window function. Furthermore, scales of loudness functions proposed by Stevens is applied to the received output [203, 204, 208]. In particular,

$$z(i, j) = \log(NSD(i, j)). \quad (4.71)$$

Finally, DCT is performed in order to achieve feature decorrelation, energy compaction, and dimensionality reduction of CFCC feature vector [33].

CFCCIF is an extension of the CFCC feature set, where the information obtained from IF is combined with the CFCC feature set. The CFCCIF feature set was first time proposed in [65] for SSD task in ASVSpooof 2015 challenge during INTERSPEECH 2015. In CFCCIF feature set, IFs were estimated using the traditional segmental Hilbert transform-based approach. The details of the Hilbert transform approach, are given in the Appendix B. However, IF can be more effectively estimated for speech signal using ESA as explained in [31]. The brief

explanation of the ESA is given in Section 4.3. Among the various DESA algorithms, DESA-1a algorithm is utilized in this study to estimate the instantaneous frequency.

In CFCC feature set, the averaging operation is performed on each subband signal while computing the NSD. It performs the lowpass filtering operation to suppress the fast temporal modulations in the subband signals [229]. Furthermore, sharp variations in the phase of the travelling wave occurs at every center frequency of the cochlear subband filter from base to the apex of the BM [230]. We argue that these sharp variations are represented by IFs. We know that vocoder-based speech signal is lacking the phase information as in natural speech signal. Furthermore, temporal discontinuities are present due to joining of speech sound units in the vocoder-dependent speech signal. To capture this distorted phase information and temporal discontinuities, we propose to use the Average Instantaneous Frequency (AIF) with the envelope representations obtained in the CFCC feature set. To that effect, let $x_i(t)$ be the subband signal corresponding to the i^{th} subband filter. Framing and windowing is performed on each subband signal, and framewise average IF is estimated for each subband as follows:

$$AIF(i, n) = \frac{1}{w} \sum_{c=m}^{m+w-1} IF(i, c), \quad m = 1, h, 2h, \dots; \forall i, n, \quad (4.72)$$

where w is the window length, n is the frame count, and h is the hop size of window function. Let $z(i, t)$ represents the combination of the NSD (in eq. (4.70)) and AIF (in eq. (4.72)) for i^{th} subband and mathematically, it is estimated as [29]:

$$z(i, t) = S(i, t) \times AIF(i, t). \quad (4.73)$$

The study reported in [231] uses the feature combination strategy by concatenating the average IFs with envelope features. However, it results in increase in dimension of the feature vector by a factor of two. Motivated from the original study in [232], the relative perceptual importance of the envelope and fine time structure is investigated by synthesizing the *auditory chimeras*; which has the envelope of the one sound and the fine structure of the other sound. Here, auditory chimeras are formed by the multiplication of the envelope and fine structure. The similar strategy is followed in CFCCIF-ESA feature set, where the multiplication of the NSD and AIF is performed. It helps to reduce the dimensionality of the feature vector. Furthermore, the random IF estimated in silence regions will be suppressed by multiplication operation as silence region possess low amplitude

values in the envelope structure. The partial derivative operation performed on $z(i, t)$ can be expressed using a chain rule, i.e.,

$$\frac{\partial z(i, t)}{\partial t} = AIF(i, t) \frac{\partial S(i, t)}{\partial t} + S(i, t) \frac{\partial AIF(i, t)}{\partial t}. \quad (4.74)$$

From eq. (4.74), it can be observed that the derivative of the $z(i, t)$ consists of two terms. The first term represents the changes in the NSD weighted by the AIF, whereas the second term represents the changes in the AIF weighted by the NSD. Furthermore, the DCT operation (as in the other cepstral features) is performed on the derivative of the $z(i, t)$, to obtain energy compact decorrelated CFCCIF feature set with a reduced dimension of feature vector.

As discussed earlier, in the CFCCIF feature set, IFs were estimated using the traditional segmental Hilbert transform-based approach, which uses the speech segment (of 10-30 ms duration) to derive the instantaneous (i.e., analytic) phase. Hence, the traditional Hilbert transform-based analytic signal generation approach requires signal for a longer duration (i.e., it is segmental approach) and hence, it averages out (blunts) the fine variations in IFs *w.r.t.* time. Whereas, in the proposed CFCCIF-ESA feature set, IFs were estimated using ESA, which uses only a few adjacent (5 – 7) samples in order to estimate the IF [194]. The ESA is not only computationally efficient but also captures instantaneously adapting nature of the modulation pattern in the time-varying speech signal [183, 194, 233]. Hence, ESA is able to estimate the IFs more accurately for the time-varying signals with the constraint of the narrowband signal (because the concept of IF is primarily developed for the monocomponent signal [234]). The functional block diagram depicting feature extraction procedure for CFCC, CFCCIF, and CFCCIF-ESA is shown in Figure 4.21. Furthermore, MATLAB pseudocode for the proposed CFCCIF-ESA feature set is illustrated in Algorithm 4.

4.6.2 Analysis of Phase-Related Artifacts in SS and VC Spoof

The phase-based information is generally neglected in several speech technology applications [235]. However, recently there are numerous amounts of evidences, which suggest the significance of the phase information. In [236], the complementary information in phase is utilized to enhance the performance of the SSD system against replay spoofing attacks. In addition, the person identification system using humming exploited the phase information for better performance [237, 238]. The phase information is also exploited in many other speech signal processing applications, such as speech enhancement, source separation, speech synthesis,

Algorithm 4 MATLAB Pseudo Code of Proposed CFCCIF-ESA Feature Set Extraction. After [1].

1. $fbankC = Cochlear_fbank(Q, \alpha, \beta, u)$, construct the cochlear filterbank
with Q subband filters and optimum set of values α , β , and u ,
 2. **for** $i = 1 : Q$ **do**,
 $y(i, :) = filter(fbankC)(i, :, 1, x)$, subband filtering using i^{th}
subband filter,
 $H_{subband}(i, :) = (y(i, :))^2$, hair cell function,
 $H_{frames} = enframe(H_{subband}(i, :), win_len, win_shift)$, framing
with appropriate window length and window shift,
 $NSD(i, :) = mean(H_{frames})$, estimation of NSD,
 $IF(i, :) = IF_using_ESA(y(i, :))$, estimation of IF using ESA,
 $Z(i, :) = NSD(i, :) \times IF(i, :)$, combination of NSD and IF,
 $F(i, :) = Z(i, n) - Z(i, n - 1)$, single sample backward difference,
end for
 3. $F_{log} = log(F)$, logarithmic operation,
 4. $F_{static} = DCT(F_{log})$, static coefficients,
 5. $F_{\Delta} = delta(F_{static})$, velocity coefficients,
 6. $F_{\Delta\Delta} = delta(F_{\Delta})$, acceleration coefficients,
 7. CFCCIF-ESA = $[F_{static}; F_{\Delta}; F_{\Delta\Delta}]$, CFCCIF-ESA feature set.
-

speech and speaker recognition [235, 239–243].

In this Section, we analyze the artifacts generated by SS and VC methods utilized in ASVSpooof 2015 challenge. SS is being extensively used in various applications, such as e-book readers, speech-to-speech translation systems, and spoken dialogue systems. The advanced approaches, such as the unit selection approach of SS, can select the appropriate speech sound units and concatenate them in ordered sequence to generate the synthesized speech, which can adopt the speaker's identity and linguistic content. Hence, SS-based spoofed speech signal can breach the ASV system. The vulnerability of the ASV system to SS-based attack was studied in [28, 244]. It has been observed that the dynamic variation of the speech signal parameters in synthesized speech signal is much lesser than the parameters of the natural speech signal [244]. It is also studied that, human auditory system is relatively less sensitive for phase spectrum than the magnitude spectrum characteristics [42, 245]. To that effect, SS-based vocoders are designed in such a way that they do not take the phase characteristics (in time and frequency-domain) into account. Hence, there exist the difference in phase characteristics of natural *vs.* synthesized speech, which can be exploited for the SSD task. To investigate this issue, the IFs of the natural and SS-based speech signals are depicted in Figure 4.24 and Figure 4.25, where IFs were estimated using ESA. IFs represents the derivative of time-domain phase characteristics. It can be observed that

the dynamic variations in IFs (indicating energy modulations in acoustic signal generation process [193]) for genuine speech signal is much more than that of the synthesized speech.

VC technique converts the given source speaker’s voice to target speaker’s voice [68, 246]. In this conversion process, input speech signal characteristics, such as voice timbre, F_0 , and duration, are mapped to that of target speaker(s). It can be achieved through the various spectral mapping techniques. In VC, vocoders are used similar to that of SS technology. Hence, artifacts introduced by the vocoders into the converted speech signal would be the similar to that of synthesized speech.

Similar analysis is performed on ASVSpooF 2019 LA dataset, where synthetic speech signal was produced by advanced neural-network-based waveform modelling techniques, known as *WaveNet* architectures [168]. These architectures can produce the synthetic speech signals as natural speech produced by the humans. However, the analysis performed to detect the artifacts using IFs shows that the difference between the genuine *vs.* spooF speech signal for ASVSpooF 2019 LA dataset is not distinct as compared to that of ASVSpooF 2015 dataset. This might be due to the fact that the nature of the IFs estimated for the synthetic speech signals in ASVSpooF 2019 dataset would be as natural as that of genuine speech signal. Thus, the effectiveness of the IF approach is determined by the vocoder. In particular, with neural-network-based vocoders, the proposed IF estimation approach does not work. Given this, the key objective of this study is to develop an effective feature set for ASVSpooF 2015 dataset that captures these variations in IFs in the traditional theoretical framework of CFCCIF, where IF is estimated using ESA.

4.6.3 Experimental Setup

The performance of the proposed CFCCIF-ESA feature set is evaluated using ASVSpooF 2015 and ASVSpooF 2017 datasets, which are described in brief in sub-Section 3.2.1 and sub-Section 3.2.2, respectively. The performance of the proposed feature set is compared against the state-of-the-art feature sets, such as CQCC, MFCC along with companion feature sets, such as CFCC and CFCCIF. For ASVSpooF 2015 dataset, GMM and CNN classifiers are utilized. The theoretical explanation of the GMM and CNN classifiers is provided in sub-Section 3.4.1 and sub-Section 3.4.3, respectively. The details of the CNN architecture utilized for the proposed CFCCIF-ESA feature set on ASVSpooF 2015 dataset is shown in Table 4.23. Whereas, for ASVSpooF 2017 dataset, GMM-based classifier is utilized.

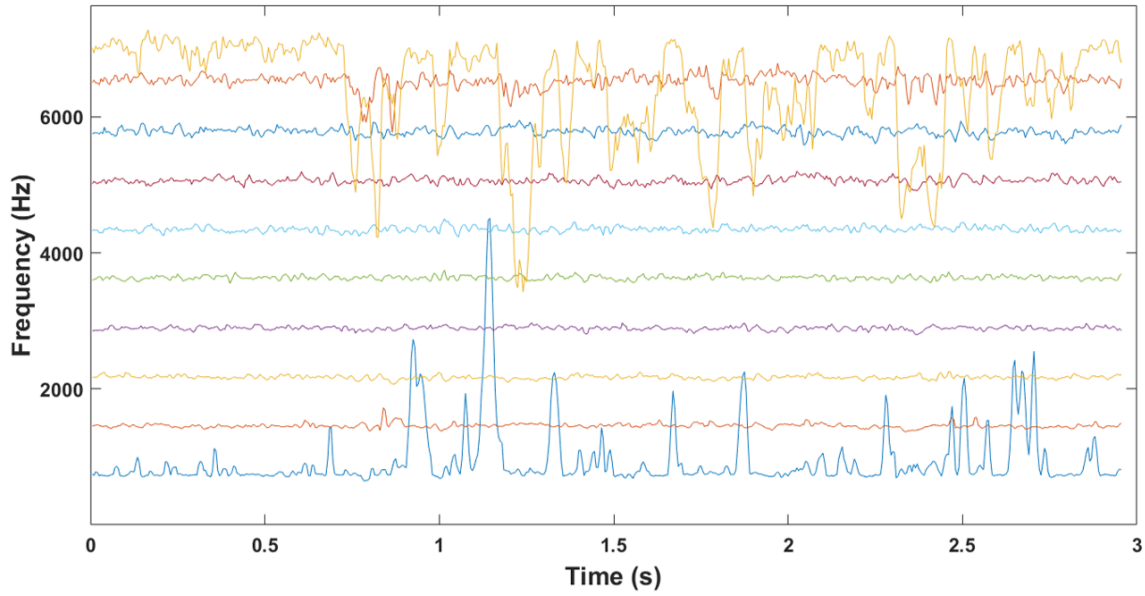


Figure 4.24: Ten IF Contours of the Subband Filtered Genuine Speech Signal. Subband Filtering is Performed by the Cochlear Filterbank with Ten Subband Filters (and hence, Ten IF Contours) Covering the Nyquist Sampling Frequency Range. After [1].

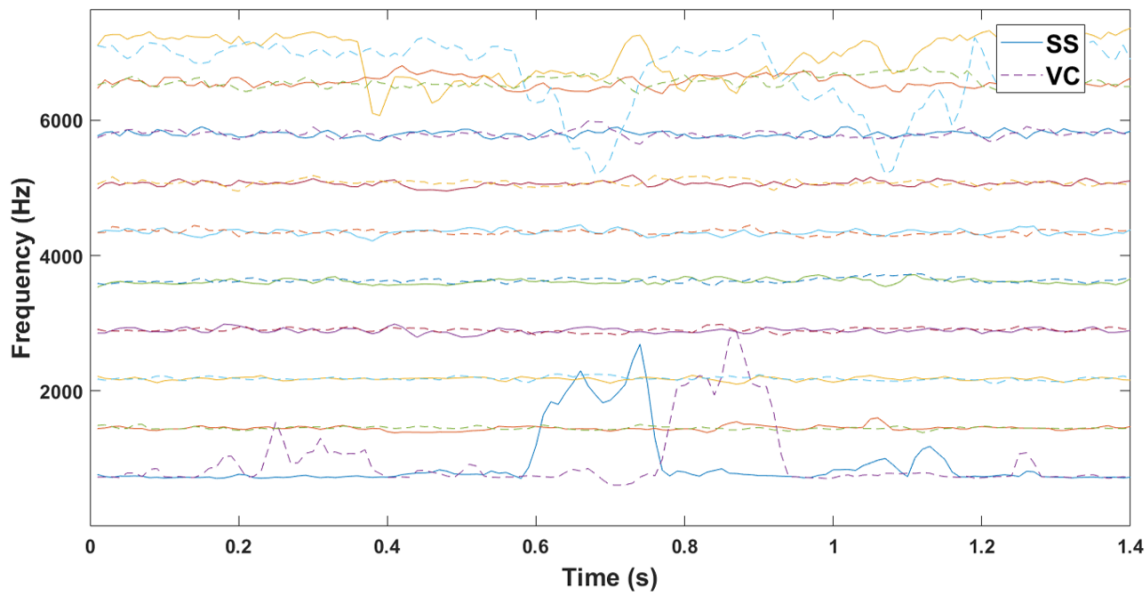


Figure 4.25: Ten IF Contours of the Subband Filtered SS- and VC-based Spoof Speech Signal. Subband Filtering is Performed by the Cochlear Filterbank with Ten Subband Filters (and hence, Ten IF Contours) Covering the Nyquist Sampling Frequency Range. After [1].

Table 4.23: Details of the Proposed CNN Architecture for SSD System. After [1].

Layer	Filter/Stride	Output	Parameters
Conv1	5 x 5/1 x 1	16 x 16 x 398	416
BN1	-	16 x 16 x 398	32
MaxPool1	2 x 2/2 x 2	16 x 8 x 199	-
Conv2	3 x 3/1 x 1	32 x 8 x 199	4640
BN2	-	32 x 8 x 199	64
MaxPool2	2 x 2/2 x 2	32 x 4 x 99	-
Conv3	3 x 3/1 x 1	64 x 4 x 99	18496
BN3	-	64 x 4 x 99	128
MaxPool3	2 x 2/2 x 2	64 x 2 x 49	-
Conv4	3 x 3/1 x 1	16 x 2 x 49	9232
BN4	-	16 x 2 x 49	32
MaxPool4	2 x 2/2 x 2	16 x 1 x 24	-
FC5	-	1 x 200	77000
FC6	-	1 x 2	402

The score-level fusion approaches, i.e., linear fusion and using logistic regression solution are employed to combine the complementary information (Chapter 3, Section 3.6) in various SSD systems.

4.6.4 Experimental Results on ASVSpooof 2015 Dataset

This Section begins with the comparison of the proposed CFCCIF-ESA feature set with the other feature sets reported in the literature on ASVSpooof 2015 challenge dataset. Furthermore, it proceeds with the detailed analysis of results obtained by the proposed CFCCIF-ESA feature set, which includes the parameter tuning and evaluation using various performance metrics.

4.6.4.1 Comparison with Other Feature Sets on Eval Set

In this Section, the performance of the proposed CFCCIF-ESA feature set is compared against the earlier reported studies in the literature on ASVSpooof 2015 challenge dataset. The Eval set consists of the *unknown attacks*, which successfully assesses the generalization capability of the SSD system. Table 4.24 shows the results on the Eval set for the SSD architectures developed on ASVSpooof 2015 challenge dataset using conventional GMM and SVM classifiers, whereas Table 4.25 shows the results on Eval set for the SSD systems, which uses DNN architectures either

for feature representation or as a classifier. The brief details of the systems shown in Table 4.24 and Table 4.25 are already discussed in Chapter 2 (Section 2.2).

From the studies reported in the literature and *w.r.t.* Table 4.25, it can be observed that the CQCC and the other feature sets derived from the CQT are producing relatively better results and thus, shows their generalization capability. The results of the proposed CFCCIF-ESA feature set trained on GMM and CNN classifiers are shown in Table 4.24 and Table 4.25, respectively. The CFCCIF-ESA-GMM system shows the significant improvement in the performance over all the other feature sets, except CQT-derived feature sets. In addition, the proposed feature set shows relatively moderate performance on known attacks, however, it shows the significant performance improvement (in particular, 3.05 % EER) on S10 attack as compared to the other feature sets except CQT-derived feature sets. Here, S10 attack uses unit selection approach for speech synthesis, and it is known to be most difficult to detect for the other feature sets. AEER for unknown attacks using CFCCIF-ESA-GMM system is as low as 1.00 %. This shows the generalization capability of the proposed feature set for the realistic unknown attack scenarios. Furthermore, when CNN is employed with CFCCIF-ESA feature set, it shows improved results over its GMM-based counterpart. The score-level fusion using linear weighted fusion of the CNN- and GMM-based SSD systems shows the % AEER reduced to 0.31 %, when fusion parameter β in eq. (3.16) is determined using the Dev set. In addition, fusion is also performed using logistic regression solution of score calibration using Bosaris toolkit, where the parameters in eq. (3.17) are trained on Dev set to obtain the best possible performance, and then used on Eval set for the score calibration. However, when fusion is performed directly onto the scores obtained on the Eval set, then it shows the AEER as 0.29 % (which is referred to as *ideal* in the last row of Table 4.25). It suggests that the GMM- and CNN-based classifiers captures the *complementary* information for the proposed CFCCIF-ESA feature set in the SSD task. In addition, it can be observed that, our proposed CFCCIF-ESA feature set shows significantly better performance for S10-attack, which is difficult to detect for the other feature sets. The performance analysis of various SSD systems, which are showing promising results on S10-attack, is discussed further in more details in Section 4.6.4.4.

Table 4.24: Results (in % EER and % AEER) from ASVSpooF Literature for Various SSD Systems Trained using GMM/SVM. The Performance of the Proposed Feature Set is Compared Against the Other Feature Sets in the Literature. After [1].

SSD System	Known Attacks					Unknown Attacks					All		
	S1	S2	S3	S4	S5	AEER	S6	S7	S8	S9	S10	AEER	AEER
CFCCIF + MFCC [29]	0.101	0.863	0.0	0.0	1.075	0.408	0.846	0.242	0.142	0.346	8.490	2.013	1.21
i-vector [75]	0.004	0.002	0.0	0.0	0.013	0.008	0.019	0.0	0.015	0.004	19.57	3.992	1.96
LFCC-DA [76]	0.027	0.408	0.0	0.0	0.114	0.110	0.149	0.011	0.074	0.027	8.185	1.670	0.89
CQCC-A [83]	0.005	0.106	0.0	0.0	0.130	0.048	0.098	0.064	1.033	0.053	1.065	0.462	0.25
CFCC [29]	0.04	1.39	0.00	0.00	2.30	0.75	1.04	0.12	0.06	0.21	12.28	2.74	1.74
CFCCIF [29]	0.03	0.72	0.00	0.00	2.24	0.60	0.98	0.16	0.88	0.29	15.42	3.55	2.07
CFCCIFS [29]	0.03	0.50	0.00	0.00	1.74	0.45	0.71	0.14	0.96	0.16	11.71	2.73	1.60
CFCCIF-ESA (A)	0.11	0.52	0	0	1.14	0.35	0.77	0.25	0.56	0.39	3.05	1.00	0.68
ISOBT [85]	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000	0.000	16.840	3.374	1.687
SFCC [85]	0.010	0.840	0.000	0.000	0.050	0.18	0.310	0.010	0.100	0.020	9.170	1.922	1.051
ISFCC [85]	0.010	0.060	0.020	0.020	0.020	0.065	0.180	0.030	0.080	0.090	8.150	1.706	0.866
FFV [89]	0.06	6.21	0.01	0.01	1.58	1.57	5.09	1.38	0.08	1.19	18.59	5.27	3.42
CAF [89]	0.00	0.02	0.00	0.00	0.00	0.004	0.10	0.02	0.00	0.02	0.30	0.44	0.05
MRP [88]	0.000	0.009	0.000	0.000	0.036	0.009	0.025	0.011	0.004	0.000	7.556	1.519	0.764
SoE [86]	0.00	0.375	0.00	0.00	0.18	0.11	0.16	0.02	0.087	0.022	15.30	3.12	1.61
SCC [84]	0.01	0.12	0.00	0.00	0.02	0.02	0.01	0.01	0.03	0.01	3.94	0.33	0.18
RPS [77]	-	-	-	-	-	0.21	-	-	-	-	-	8.883	4.547
RPS + MGDCC [78]	0.00	0.009	0.00	0.00	0.015	0.005	0.081	0.005	0.080	0.00	37.06	7.44	3.72
M&P feats. * [81]	0.024	0.104	0.025	0.016	0.032	0.041	0.093	0.010	0.236	0.000	26.392	5.347	2.694
M&P feats. * [79]	0.173	0.610	0.319	0.289	0.399	0.358	0.906	0.242	0.417	0.246	28.581	6.078	3.218

* M&P feats. is acronymed for fusion of the various magnitude- and phase-based features.

Table 4.25: Results (in % EER and % AEER) from ASVSpooof Literature for Various SSD Systems Trained using DNN architectures. The Performance of the Proposed Feature Set is Compared Against the Other Feature Sets in the Literature. After [1].

SSD System	Known Attacks						Unknown Attacks						All	
	S1	S2	S3	S4	S5	AEER	S6	S7	S8	S9	S10	AEER	AEER	AEER
s-vector [94]	-	-	-	-	-	0.058	-	-	-	-	-	4.998	2.528	
M & P feats. [90]	0.00	0.00	0.00	0.00	0.01	0.002	0.01	0.00	0.00	0.00	26.1	5.22	2.62	
Spectro/CNN [91]-a	0.08	0.19	0.02	0.03	1.26	0.31	1.48	0.68	0.01	0.16	26.83	5.83	3.07	
Spectro/RNN [91]-b	1.21	0.79	0.24	0.39	1.77	0.87	0.87	0.96	0.04	0.41	17.97	4.05	2.46	
Spectro/CNN + RNN [91]-c	0.16	0.50	0.03	0.03	1.38	0.40	0.85	0.91	0.03	0.59	14.27	3.33	1.86	
Fusion [91]-(a+b+c)	0.09	0.29	0.00	0.00	0.99	0.27	0.64	0.71	0.00	0.29	11.67	2.66	1.47	
(Spectrum + RPS)/ DNN [92]	0.021	0.031	0.021	0.023	0.031	0.025	0.038	0.032	0.041	0.021	40.708	8.168	4.096	
CQSPIC-A [93]	0.00	0.00	0.00	0.00	0.004	0.00	0.00	0.00	0.008	0.00	0.368	0.075	0.038	
CQUEST-DA [95]	0.00	0.00	0.00	0.00	0.009	0.00	0.005	0.004	0.089	0.00	0.456	0.110	0.056	
CMC-A [96]	0.00	0.00	0.00	0.00	0.004	0.00	0.005	0.00	0.026	0.00	0.221	0.050	0.026	
CFCCIF-ESA-CNN (B)	0.03	0.15	0.01	0.01	0.20	0.08	0.17	0.08	3.88	0.19	1.15	1.09	0.58	
(A) + (B) *	0.03	0.10	0.00	0.00	0.23	0.072	0.18	0.05	1.25	0.13	1.13	0.548	0.31	
(A) + (B) +	0.03	0.07	0.00	0.00	0.20	0.06	0.15	0.04	1.35	0.08	1.45	0.614	0.337	
(A) + (B) ideal	0.03	0.01	0.00	0.00	0.25	0.058	0.19	0.05	1.06	0.13	1.20	0.52	0.29	

‘*’ denotes the weighted linear fusion, whereas, ‘+’ denotes the fusion using Bosaris toolkit, which provides logistic regression solution for score calibration. M&P feats. is acronymed for fusion of the various magnitude- and phase-based features.

4.6.4.2 Detailed Analysis

- Parameter Tuning

The results reported in Table 4.24 and Table 4.25 for the proposed CFCCIF-ESA feature set are obtained by performing parameter tuning (such as feature parameters α and β , number of subband filters in the filterbank, and dimension of feature vector) on Dev set with GMM classifier. As given in [29], we performed the initial experiment with $\beta = 0.035$ and varying α with 40 subband filters in the filterbank. In addition, the feature vector dimension is set to 36, which includes 12-static, Δ , and $\Delta\Delta$ features. It was observed that $\alpha = 3$ gives relatively better performance. Then, β is varied by keeping the value of $\alpha = 3$. Better performance was obtained with $\alpha = 3$, and $\beta = 0.005$, which determines the *shape* and *width* of the frequency response, respectively, of the subband filters in the cochlear filterbank. Then, by fixing the optimized value of α and β , the experiments are performed with varying number of subband filters in the cochlear filterbank. It can be observed from Figure 4.26 that the better results are obtained as we keep on increasing the number of subband filters in the filterbank. It has been observed that, the performance is improved as the number of subband filters reaches to 60, and then very little variation in the % EER was observed after 60 subband filters. It might be due to the fact that TEO is historically developed for monocomponent signals (as discussed in Section 4.3), and modelling the energy of SHM [182]. However, the bounds derived in [194] suffices the applicability of TEO for subband filtered signals for estimating the reasonably accurate IFs. In addition, human auditory system for hearing depends upon thousands of subband filters [40]. Furthermore, with increase in number of subband filters in the cochlear filterbank, the approximation errors may keep on reducing and hence, the estimated IFs have the more accurate values with increase in subband filters. It is obvious that with the greater number of subband filters, each subband filtered signal is approximating the monocomponent nature of the signal, and helps to predict IFs more accurately. Furthermore, experiments are performed using the varying number of dimensions of the CFCCIF-ESA feature vector. It has been observed that relatively better performance is obtained with 18-D CFCCIF-ESA feature set, which includes static, Δ , and $\Delta\Delta$ features.

- Effect of the Various Filterbanks

The proposed feature set adopts a cochlear filterbank using linearly-spaced subband filters. To validate the effectiveness of the cochlear filterbank in CFCCIF-ESA feature set, experiments are performed with the other filterbanks, such as

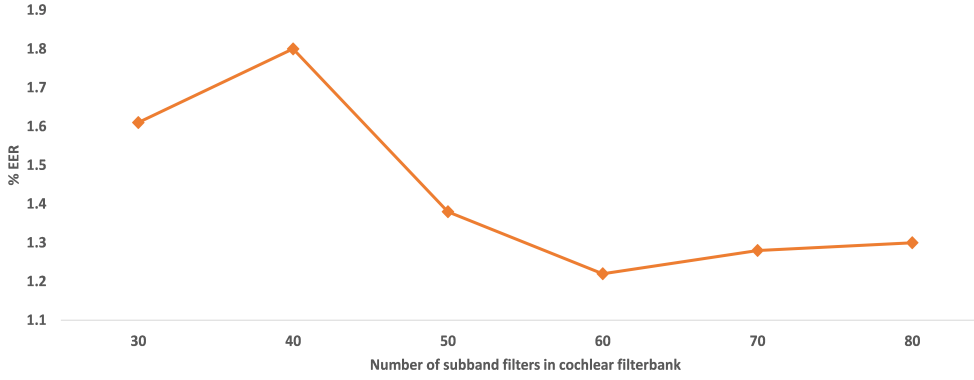


Figure 4.26: Results (in % EER) *w.r.t.* Number of Subband Filters on the Dev Set. After [1].

Gammatone and Gabor filterbanks, which allows to derive subband-based features for speech signal processing applications [41,247]. The cochlear filterbank is replaced by Gabor and Gammatone filterbanks by keeping the remaining architecture intact as that of the proposed feature set. Table 4.26 shows the results for the proposed CFCCIF-ESA feature set and the feature sets derived by replacing the cochlear filterbank with Gabor and Gammatone filterbanks in the proposed feature set architecture. The impulse response of the Gabor’s subband filter is given by [31]:

$$g(t) = \exp(-a^2t^2) \cdot \cos(\omega_c t), \quad (4.75)$$

where ω_c represents center frequency of a Gabor filter, and a controls its bandwidth. From Table 4.26, it can be observed that the performance of the cochlear filterbank shows the significant improvement over Gabor and Gammatone filterbanks. Moreover, the performance of the Gabor filterbank is relatively nearer to the cochlear filterbank, which is quite expected due to their similar mathematical structure as can be seen from eq. (4.65) and eq. (4.75).

To analyze the effect of filterbank on the performance of the proposed CFCCIF-ESA feature set, we have observed the frequency response of various filterbank structures consisting of 10 subband filters and occupying the entire frequency range for a given sampling frequency. From Figure 4.27, it can be observed that the bandwidth of the cochlear filterbank is narrowest (and hence, having relatively best quality factor) among the three filterbanks. From the time-frequency analysis literature [248], IF of the monocomponent signal is defined as the frequency of the sinusoid, which locally fits into infinitesimal smaller window applied onto the signal and hence, IF estimation is expected to be relatively accurate for narrowband filtered output signal than its monocomponent counterpart. The study reported in [31] shows that IFs can be more accurately estimated for the nar-

rowband signals. From Figure 4.27, it can be observed that the subband filtered signals obtained from the cochlear filterbank would be the more narrowband as compared to that of Gabor and Gammatone filterbanks. We know that,

$$Q \propto \frac{1}{B}, \quad (4.76)$$

where Q and B represents *quality factor* and *bandwidth*, respectively, of the subband filters. Thus, cochlear filterbank will have higher quality factor and higher frequency selectivity. Hence, the proposed feature set framework, which utilizes IFs, might be more suitable using cochlear filterbank. Moreover, the subband filters of the Gammatone filters in higher frequency range, have higher bandwidth (and thus, poor quality factor and hence, poor frequency selectivity). And hence, IFs estimated from the subband filtered signals using Gammatone filterbank would not be accurate. This might be the reason that the Gammatone filterbank is producing relatively bad performance in the CFCCIF-ESA framework.

Table 4.26: Results (in % EER) on Proposed CFCCIF-ESA Feature Set Framework with Various Filterbank Structures. After [1].

Filterbank Structure	Dev	Eval
Gabor	1.03	2.56
Gammatone	13.13	10.56
Cochlear	0.6	0.85

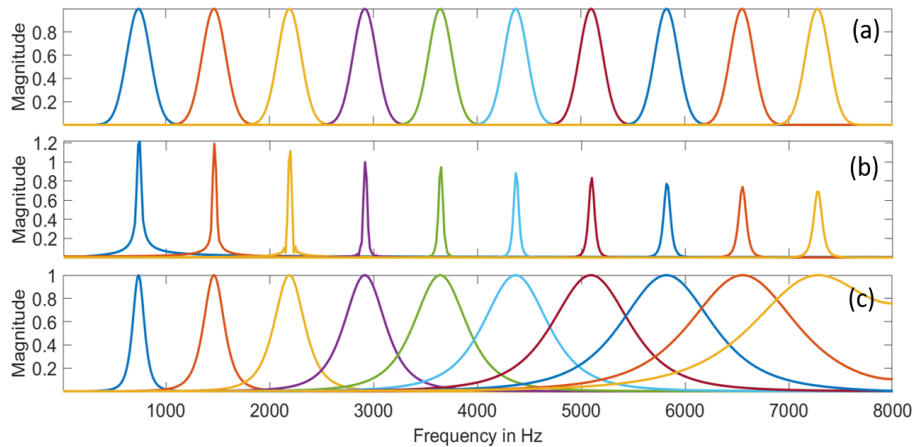


Figure 4.27: Frequency Response of Filterbanks with 10 Subband Filters: (a) Gabor, (b) Cochlear ($\alpha = 3$ and $\beta = 0.005$), and (c) Gammatone. After [1].

Table 4.27: Results in % EER, % Classification Accuracy, and AUC of Intersection of the Probability Density Functions (*pdfs*) Obtained from the LLR Scores for Dev and Eval Set of ASVSpooF 2015 Dataset. After [1].

SSD System	% EER		AUC		% Classification Accuracy	
	Dev	Eval	Dev	Eval	Dev	Eval
CFCC-GMM	1.47	1.77	0.0368	0.0560	98.19	96.64
CFCCIF-GMM	0.78	1.27	0.0266	0.0433	98.43	96.63
CFCCIF-ESA-GMM (A)	0.6	0.85	0.0204	0.0390	98.52	97.23
CFCCIF-ESA-CNN (B)	0.044	0.82	0.0037	0.0328	99.73	97.45
(A) + (B)	0.028	0.45	0.0031	0.0206	99.69	98.38

4.6.4.3 Assessment of the Proposed CFCCIF-ESA Feature Set using Various Performance Metrics

In this study, we assessed the performance of the proposed CFCCIF-ESA feature set using three performance metrics, namely, % EER, % classification accuracy, and AUC of the overlapping regions for the *pdfs* of the LLR scores for genuine and spoof speech utterances. Table 4.27 shows the performance of the proposed CFCCIF-ESA feature set along with the earlier cochlear filter-based features, namely, CFCC and CFCCIF². The feature parameters of the CFCC, CFCCIF, and CFCCIF-ESA are set for corresponding best possible results. It can be observed from Table 4.27 that the proposed feature set gives better % classification accuracy and less AUC than the CFCC and CFCCIF feature sets. In addition, results are improved further using CNN. Furthermore, the classifier-level fusion of the GMM and CNN using proposed feature set gives all the more better results.

Figure 4.28 shows the LLR score densities of genuine *vs.* spoof speech utterances obtained from the SSD systems with various feature sets and classifiers. Figure 4.28 (a), (b), (c), (d), and (e) shows the genuine *vs.* spoof speech LLR score distribution on Dev set obtained from CFCC-GMM, CFCCIF-GMM, CFCCIF-ESA-GMM (A), CFCCIF-GMM-CNN (B), and score-level fusion of the SSD system (A) and (B), respectively. However, Figure 4.28 (f), (g), (h), (i), and (j) shows the score distribution on the Eval set obtained from CFCC-GMM, CFCCIF-GMM, CFCCIF-ESA-GMM (A), CFCCIF-GMM-CNN (B), and score-level fusion of the SSD systems (A) and (B), respectively. It can be observed from Table 4.27 that the AUC for the overlapping region between the *pdfs* of the LLR scores for genuine and

²Here, feature parameters of CFCC feature set shows the better performance for the parameter set as suggested in [29] (i.e., $\alpha=3$, and $\beta = 0.035$) with 60 subband filters in the filterbank. However, CFCCIF and CFCCIF-ESA feature sets are tuned with the parameter set as explained in sub-Section 4.6.4.2 (i.e., $\alpha=3$, and $\beta = 0.005$, 60 subband filters in the filterbank and 18-dimensional feature set, which includes static, Δ , and $\Delta\Delta$ features).

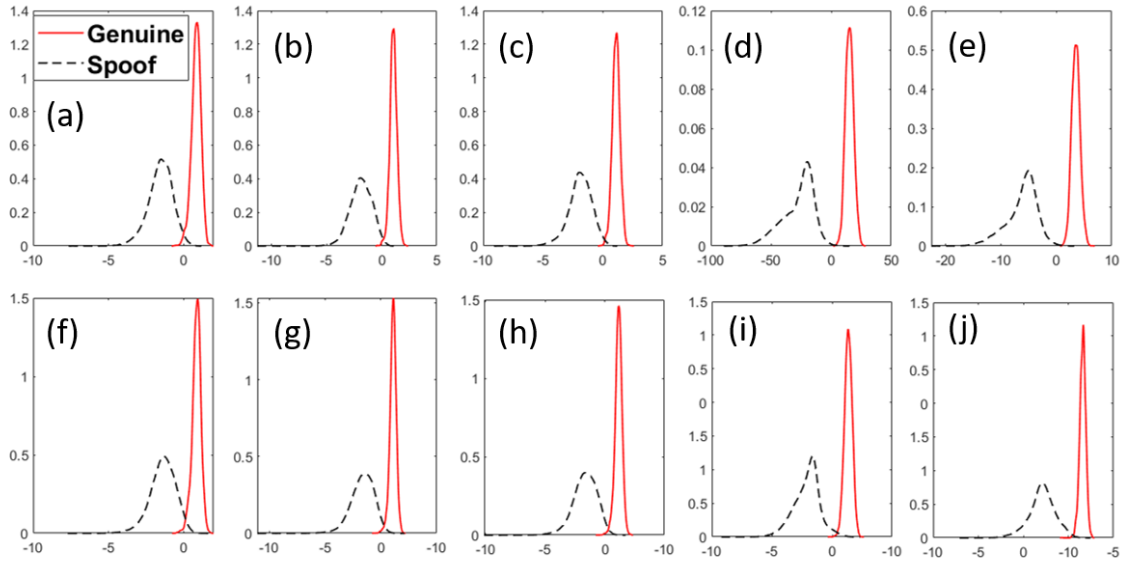


Figure 4.28: LLR Score Distribution of Genuine *vs.* Spoof Speech Distribution for the SSD Systems Developed using Various Feature Sets and Classifiers. (Figure 4.28 (a) and (f)), (Figure 4.28 (b) and (g)), (Figure 4.28 (c) and (h)), (Figure 4.28 (d) and (i)), and (Figure 4.28 (e) and (j)) Shows the LLR Score Distribution for the SSD Systems CFCC-GMM, CFCCIF-GMM, CFCCIF-ESA-GMM (A), CFCCIF-ESA-CNN (B), Score-Level Fusion of (A) and (B) on Dev and Eval Set, Respectively. After [1].

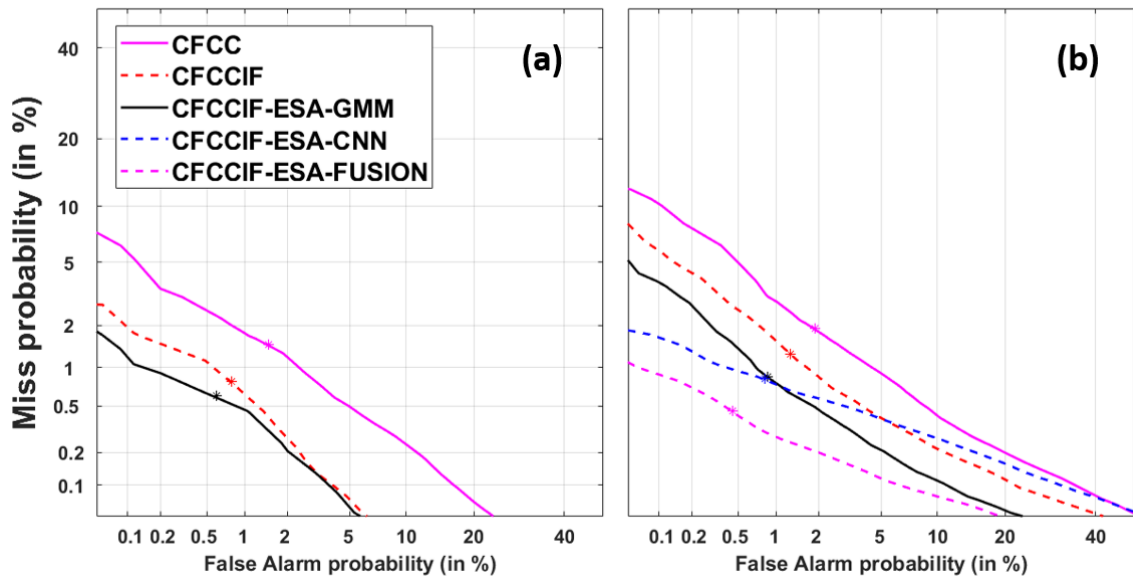


Figure 4.29: DET Curves Obtained from the SSD Systems Implemented using CFCC, CFCCIF, and Proposed CFCCIF-ESA Feature Sets on ASVspoof 2015 dataset. Figure 4.29(a) and Figure 4.29(b) Shows the DET Plots for Dev and Eval Set, Respectively. DET Curves for CFCCIF-ESA-CNN and CFCCIF-ESA-Fusion are not Visible in Figure 4.29(a) Due to % EER is Approaching to Zero. After [1].

spoof speech utterances is *minimum* for the proposed CFCCIF-ESA feature set. The AUC is larger for the CFCC feature set than the CFCCIF/CFCCIF-ESA feature sets, which indicates the significance of the IF in the proposed feature set architecture for relatively better performance. The exact values of the AUC are shown in Table 4.27. Furthermore, score-level fusion of the CFCCIF-ESA-GMM and CFCCIF-ESA-CNN shows the significant reduction in the AUC.

Figure 4.29 (a) and Figure 4.29 (b) shows the DET curves for the CFCC-GMM, CFCCIF-GMM, CFCCIF-ESA-GMM (A), CFCCIF-GMM-CNN (B), and score-level fusion of the SSD system (A) and (B) on the Dev and Eval set, respectively. As the % EER for CFCCIF-GMM-CNN system is near to zero for Dev set, we cannot observe the corresponding DET curves. The similar inferences can be drawn from DET curves as that of the *pdf* of LLR scores of the genuine *vs.* spoof speech utterances.

4.6.4.4 Performance Analysis on S10 Spoofing Attack

In ASVspoof 2015 challenge dataset, ten different kinds of spoofing attacks, labelled as S1 to S10, are utilized. To the best of author’s knowledge and belief, Table 4.28 represents the performance of the various state-of-the-art feature sets in the literature and in this study, which shows the promising performance against the S10 spoofing attack.

Table 4.28: Results (in % EER) for the Various Feature Sets for S10 Spoofing Attack Detection. After [1].

Feature Set	Classifier	% EER
CQCC-A [83]	GMM	1.065
CAF [89]	GMM	0.30
FFV-SD [89]	GMM	18.59
APGDF-A [89]	GMM	6.40
SCC [84]	GMM	3.94
CQSPIC-A [93]	DNN	0.368
CQUEST-DA [95]	DNN	0.456
CMC-A [96]	DNN	0.221
CFCCIF [29]	GMM	15.42
CFCCIFS [29]	GMM	11.71
CFCC [29]	GMM	12.28
CFCCIF-ESA	GMM	3.05
(Proposed)	CNN	1.15

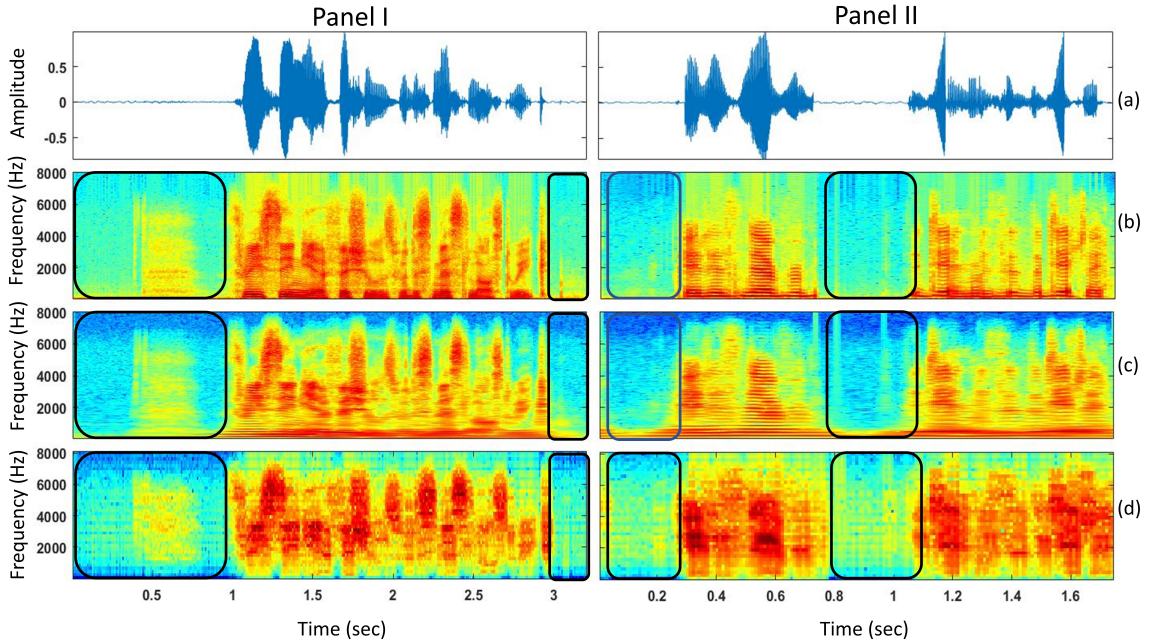


Figure 4.30: The Speech Signal and Spectrographic Representation of the Genuine *vs.* S10-attack. Panel-I and Panel-II Shows the Speech Signal with its Spectrographic Representation of the Genuine and S10-attack, Respectively. Figure 4.30(a) Represents the Speech Signal. Whereas, Figure 4.30(b), Figure 4.30(c), and Figure 4.30(d) Represents the Corresponding Spectrogram Obtained from STFT, CQT, and CFCCIF-ESA Feature Set, Respectively. After [1].

Among the feature sets reported in Table 4.28, CQCC-A, CQSPIC-A, CQUEST-DA, and CMC-A are derived from CQT, which uses the variable spectro-temporal resolution. The SSD system for CQCC-A is developed using GMM classifier, whereas CQSPIC-A, CQUEST-DA, and CMC-A uses the DNN-based classifier. All these CQT-based SSD systems performs better than our proposed SSD systems based on CFCCIF-ESA feature set. The CAF feature set is also performing better than our proposed CFCCIF-ESA feature set. However, CAF feature set is formed by concatenating the three feature sets, namely, CQCC, APGDF, and FVV. On the other hand, proposed CFCCIF-ESA feature set performs significantly better than the standalone APGDF-A and FVV feature sets, as shown in Table 4.28. The proposed CFCCIF-ESA feature set also performs relatively better than SCC feature set. The comparison is also shown with the other cochlear filter-based features, such as CFCC, CFCCIF, and CFCCIFS. Whereas, the proposed CFCCIF-ESA feature set performs significantly better than their CFCC, CFCCIF, and CFCCIFS counterparts. Overall, the proposed CFCCIF-ESA feature set performs significantly better than the other state-of-the-art feature sets, except CQT-based feature sets for S10-based spoofing attack detection.

To analyze the ability of the CQT-based feature sets *vs.* cochlear filterbank-

based feature sets for the SSD task, we observed the spectrograms for STFT, CQT, and CFCCIF-ESA feature set. Here, spectrogram of CFCCIF-ESA feature set refers to the spectral representation of CFCCIF-ESA feature set, which is obtained before DCT operation in CFCCIF-ESA feature extraction framework. Panel-I and Panel-II in Figure 4.30 shows the speech signal along with its spectrographic representations for the genuine and S10-attack, respectively. Figure 4.30(a) represents the speech signals, whereas, Figure 4.30(b), Figure 4.30(c), and Figure 4.30(d) represents the spectrogram obtained from STFT, CQT, and CFCCIF-ESA feature sets, respectively. It can be observed that the silence region is more prominently emphasized by CQT-based spectrogram than the STFT- and CFCCIF-ESA-based spectrograms. In the other words, CQT-based features are able to detect the discontinuity in energy flow of speech wave, which is very much prevalent in S10-attack, as it involves concatenation of natural speech sound units and thus, there will be a discontinuity at the joint location of speech sound units. This might be the reason for better performance of the CQT than the CFCCIF-ESA feature set [1].

4.7 Chapter Summary

In this chapter, three feature sets, namely, ETECC, CTECC, and CFCCIF-ESA are developed for the SSD task. These feature sets are derived based on the concept of TEO.

ETECC feature set is developed using the concept of ETEO, which uses the concept of signal mass to get a more precise estimate of signal energy in comparison with TEO. Particularly, the TEO-related approximation $\sin(\omega) \approx \omega$ holds true only for lower frequencies and hence, is not suitable for higher frequency contents of signals. The concept of signal mass in ETEO compensates the energy in the high frequency regions to provide a more precise estimate of signal energy. Subband filtering was performed using Gabor filterbank with linearly-spaced frequency responses. Subband filtering helps to approximate the subband filtered signal to a monocomponent signal, which eases the accurate estimation of the energies. Furthermore, PFE analysis on ASVSpooof 2017 dataset is also performed for the feature sets in this study. The extensive set of experiments are performed for parameter tuning of the proposed ETECC feature set. Furthermore, the experiments are extended to compare the performance of the state-of-the-art feature sets. The relatively better performance of ETECC is observed than the other features on ASVSpooof 2017 and ReMASC datasets.

In CTECC_{max} feature set, the multi-channel information in microphone array is

exploited for the replay SSD in VAs. To that effect, we provide the mathematical analysis for choosing the appropriate subband channel information (in particular, maximum noise distortion including acoustic reverberation due to replay attack) among the multiple subband channels obtained from the microphone array. The appropriate subband channel information is based on *maximum* cross-Teager energy (as opposed to minimum cross-Teager energy as in the speech recognition literature) estimation among the subband channels, to derive the proposed CTECC_{max} feature set. The experiments are performed using ReMASC dataset. In replay SSD, it is necessary to emphasize the acoustic effects and hence, we chose *maximum* cross-Teager energy to extract these acoustic effects. The proposed CTECC_{max} feature set outperforms the results reported in recently proposed complex deep learning-based architecture and other state-of-the-art feature sets commonly used in the anti-spoofing literature. One of the limitations of ReMASC dataset is absence of well known data partition that is universally accepted (for example, we followed data partition *w.r.t* study reported in [12]) and then, there is need to address this in the near future.

Further, the CFCCIF-ESA feature set is developed, which effectively combines the magnitude and phase information to detect the SS-, VC-, and replay spoofing attacks. The performance of the CFCCIF-ESA feature set is evaluated on ASVSpooF-2015 and -2017 datasets. The discriminative acoustic cue for SS- and VC-based attacks lies in the presence of the artifacts in synthesized and voice-converted speech signals, wherein the speech signal is generated using only magnitude information of the spectrum, neglecting the phase component during signal reconstruction. Thus, phase information in those speech signal is not as natural as in genuine speech signals. This fact is analyzed by visualizing the IFs from genuine and synthetic spoof speech signals. The proposed CFCCIF-ESA feature set combines the implicit information from magnitude envelopes and IFs estimated using ESA, from the subband filtered signals. The cochlear filterbank is utilized in the subband filtering. In this work, IFs are estimated using ESA, which have relatively low computational complexity, high time resolution, and instantaneously adapting nature, as compared to the Hilbert transformed-based approach that has poor time resolution, and requires the computationally complex task of phase unwrapping.

The capability of the ESA is reflected into better performance for SSD task. With the proposed feature set, we have employed two classifiers, namely, GMM and CNN. Parameters of the proposed feature set are fine-tuned by observing the performance of the SSD system trained using GMM, and tested on the Dev

set. In addition, the experiments are performed by replacing the cochlear filterbank by Gabor and Gammatone filterbank structures in the proposed feature set framework. Relatively better performance is observed for cochlear filterbank, which indicates that the cochlear filterbank is the better choice for given feature set framework. The estimated set of parameters for the better performance are further utilized to test the performance on the Eval set. Furthermore, it can be observed that, our CFCCIF-ESA shows significantly better performance for S10-attack, which is known to be most difficult to detect for the other feature sets reported in the anti-spoofing literature. It is observed that the CFCCIF-ESA outperforms the other feature sets except CQT-derived feature sets. Given this limitation, there is further scope for the improvement and which we believe is an open research problem. The proposed feature set can be modified *w.r.t.* the filterbank structure. In addition, the simulation of unidirectional nature of the basilar membrane (BM) movement can be thoughtfully altered by the other suitable function, while deriving the CFCC feature set. Furthermore, there is no known study analyzing the benefit of cochlear filter for speech synthesis and hence, its an open research problem.

In the next chapter, the capability of the CQT to emphasize the lower frequency region is utilized to extract the acoustic characteristics of the pop noise, which is low frequency in nature. The pop noise can be considered as the characteristics of the live speaker in front of the ASV system and hence, it can be utilized in SSD task.

CHAPTER 5

Spectral-Based Features for Anti-spoofing

5.1 Introduction

In¹ the earlier chapter, the development of the CM system against spoofing attacks using TEO-based feature sets, which are derived using subband filtering, are discussed. In this chapter, the spectral-based representations, namely, CQT and SRCC feature sets for SSD task are discussed. The CQT is utilized for VLD, where presence of the pop noise is supposed to be the characteristics of the live speech. The capability of the CQT to capture the low frequency contents is exploited for locating the pop noise characteristics. Other spectral-based SRCC feature set is proposed, which uses *power-law* nonlinearity instead of the *logarithmic* nonlinearity. The power-law nonlinearity is more desirable for feature representation as it provides the flexibility for more compressed representation. The details of the fea-

¹This Chapter is based on the following publications:

- Kuldeep Khoria, **Ankur T. Patil**, and Hemant A. Patil, "On Significance of Constant-Q Transform for Pop Noise Detection" in Computer, Speech & Language, Elsevier, vol. 77 (2023), pp. 101421.
- **Ankur T. Patil**, Kuldeep Khoria, Hemant A. Patil, "Voice Liveness Detection using Constant-Q Transform-Based Features", to appear in European Signal Processing Conference (EUSIPCO)-2022, Belgrade, Serbia, August 2022.
- **Ankur T. Patil**, Harsh Kotta, Rajul Acharya, and Hemant A. Patil, "Spectral Root Features for Replay Spoof Detection in Voice Assistants" in International Conference on Speech and Computer (SPECOM), St. Petersburg, Russia, Sept. 2021, pp. 504-515.
- Kuldeep Khoria, **Ankur T. Patil**, and Hemant A. Patil, "Significance of Constant-Q Transform for VoiceLiveness Detection" in European Signal Processing Conference (EUSIPCO), Dublin, Ireland, August 2021, pp. 126-130.
- Prasad A. Tapkir, **Ankur T. Patil**, Neil Shah, and Hemant A. Patil, "Novel spectral root cepstral features for replay spoof detection." in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, Hawaii, USA, November 2018, pp. 1945-1950.

ture set development, experimental setup, and results for CQT and SRCC feature sets are explained in subsequent Sections.

5.2 CQT for Voice Liveness Detection (VLD)

5.2.1 Literature in Brief

In the earlier chapters, the development of the CM system against spoofing attacks for ASV and VAs are discussed. In the development of those CMs, distortions are introduced by the spoof generation mechanism as an acoustic signature, which are utilized to detect the spoofing attack. However, less attention is being given towards the liveness detection of voice to avoid the possible spoofing attacks. If the distance between the speaker and microphone is less, then the microphone can capture the pop noise as an important acoustic signature for live speaker [8]. Pop noise is nothing but the distortion in the speech signal introduced by the speaker's breath sound [148]. Thus, pop noise can be attributed to the presence of live speaker, and it can be exploited for liveness detection to alleviate the spoofing attacks. To that effect, recently POCO dataset is developed, which can be used to build the countermeasure strategies against spoofing attacks by identifying the presence of pop noise present in live, i.e., genuine speaker's voice [8].

To the best of author's knowledge and belief, the problem of VLD was introduced first time in [34], where possible methodologies of the pop noise (liveness) detection were discussed. They proposed two approaches, namely, *low frequency-based single channel detection*, and *subtraction-based pop noise detection with two channels*. In the first approach, the presence of the pop noise at low frequency region is exploited to detect the liveness in the speech signal. Here, the low frequency region in the spectrogram is processed to extract the period of the pop noise in the input speech signal. The latter approach is based on the multi-channel microphone, which extracts the evidences of pop noise from the entire frequency range. However, this approach cannot succeed if the imposter embeds the pop noise from his/her own breathing. In addition, performance of these approaches depends upon microphone quality and linguistic content of the utterances. To that effect, phoneme-based pop noise detection is performed in [147], where pop noise duration is detected in the utterance and estimated phonemes in this duration are analyzed for VLD. This approach is further extended with GFCC feature set for pop noise detection in [148].

Recently, POCO is developed, which can be used to build the VLD system by identifying the pop noise, which is the characteristics of the live (genuine) speech [8]. Identifying the pop noise for live speaker detection might be useful strategy in the applications, where the testing microphone is placed at a short distance from the speaker, and consequently, this strategy may protect the ASV system from the spoofing attacks (of course, with the assumption that spoofed speech is not recorded with wiretapping). The architecture proposed in [34] is the popular approach for the VLD and consequently, it is utilized in the original POCO dataset reported in [8]. In this thesis, we exploited the CQT for VLD. The historical evolution of the CQT through STFT can be studied from [14, 202, 249–253]. The CQT and its derivatives were successfully utilized for anti-spoofing task. The timeline for the CQT and its derived feature sets for anti-spoofing task, is shown in Figure 5.1, and explained in brief as follows [15]:

- Motivated from the original work by Wiener [249] for estimation of the power spectrum, Schroeder and Atal defined the STFT for a practical and well-behaved signal (such as speech wave) [250] as opposed to work of Gabor [202], where STFT involved integration from $-\infty$ to $+\infty$.
- In [251], short-time spectral analysis is performed with non-uniform sub-band filters, which leads to invertible integral transform via Mellin transform, of course with the assumption of causal window function in STFT. In [254], the window is not restricted to be causal, however, the window argument is chosen to be the product of time and frequency and reciprocal of a factor related to Q (i.e., filter selectivity or its quality factor), where synthesis (inverse) integral involved Hilbert transform.
- In [252], the constant- Q spectral analysis was indicated as a means for implementing ‘Fourier-Mellin’ transform for speech analysis.
- Furthermore, inverting integral transforms is presented in [253], for the *short-time* and the *average-power spectrum* resulting from any constant- Q spectral analysis.
- The CQT was proposed for the first time (1991) to model the geometrical spacing between the western musical notes [14].
- In 2016, CQCC feature set was proposed as state-of-the-art feature set for anti-spoofing task on ASVSpooF 2015 dataset [83]. Recently, CQCC is also applied for infant cry classification task [26].

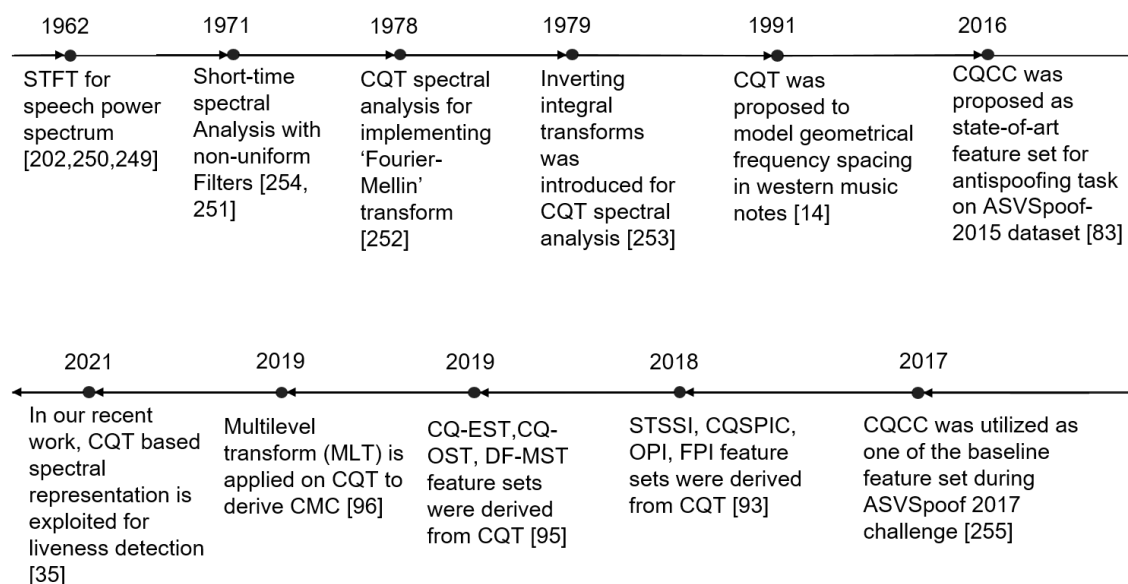


Figure 5.1: Selected Chronological Progress for CQT and its Derived Feature Sets for Speech Analysis and Anti-spoofing Tasks. After [15].

- In 2017, CQCC was utilized as one of the baseline feature set during the ASVSpooF 2017 challenge [255].
- In 2018, the feature sets derived from CQT, namely, STSSI, CQSPIC, OPI, and FPI are proposed, and these feature sets are trained using DNN classifier. Among these features, CQSPIC-A shows the remarkable performance, where -A refers to $\Delta\Delta$ -features [93].
- In 2019, other study in [95] proposes a subband transform rather than the fullband transform on CQT with three different scales, i.e., linear, octave, and Mel scale to derive three feature sets, namely, CQ-EST, CQ-OST, and DF-MST. The CQ-EST-DA (-DA refers to combination of Δ and $\Delta\Delta$ features) feature set with DNN classifier gave the better performance.
- Also, the MLT is applied on CQT to derive CMC [96].
- In our recent work, we exploited the geometric frequency spacing of the CQT-based spectral representation for the VLD task [35].

The key motivation of using CQT is its high frequency resolution in low frequency regions and hence, it is capable of capturing the prominent acoustic cues related to pop noise for VLD task. The work reported in this chapter comprise the following contributions: [35,151]

- the detailed mathematical and spectrographic analysis of CQT *vs.* STFT is presented, which demonstrates the capability of CQT over STFT to capture the details of the pop noise;
- as per Heisenberg’s uncertainty principle in signal processing framework [200], analysis of various window functions is performed in CQT *w.r.t.* window length, window type, and Heisenberg’s box;
- the performance of the VLD system for the various frequency ranges, and speaker-microphone distances is analyzed;
- a new database from POCO database is generated by simulation of replay mechanism to analyze the effect of reverberation along with pop noise;
- we extended experiments by considering two state-of-the-art deep learning architectures, namely, CNN, LCNN, and ResNet, to work as classifiers in conjunction with the proposed CQT-based features, and reported the results for VLD task;
- The experiments using proposed CQT-based feature sets are extended on ASVSpooof challenge datasets, namely, ASVSpooof 2019 PA and ASVSpooof 2017 version 2.0 dataset.

5.2.2 VLD-ASV System and Baseline

In the practice, we would expect an ASV system to be robust against any or all of the possible spoofing attacks. Replay, unlike any other spoofing attack, is the most accessible kind of spoofing attack, wherein the attacker tries to imitate the target speaker simply by replaying the pre-recorded voice samples and thus, the attacker need not have a detailed technical skills/knowledge. The replay speech recorded with a high quality recorder and playback device in a clean recording environment is very hard to detect as it is very similar to the genuine speech [256]. Hence, replay attacks are very easy to mount, however, the present ASV systems find it very hard to detect it. To that effect, we propose an efficient approach of VLD, which aims to detect presence of ‘live’ speaker in front of ASV system. In VLD, it uses the pop noise as an acoustic signature of the live speaker and it can assist the current countermeasure strategies of anti-spoofing for ASV. Detailed discussion of pop noise, VLD, STFT-based baseline, and proposed CQT-based feature extraction is discussed in the next sub-Sections.

5.2.2.1 Pop Noise and VLD System

During natural speech production mechanism, airflow travels from the lungs to the vocal folds. This airflow excites the vocal tract system, which can be modeled as the cascade of several 2^{nd} order resonators (i.e., the organ pipes). This model represents the bursts of air coming out of the organ pipe (i.e., mouth and nostrils) [42]. If this sound wave is captured at a short distance from the microphone and a speaker, the microphone along with the speech signal also captures the friction between the lips as *bursts*. The sound produced because of these *bursts* is termed as *pop noise* [34]. The speaker-microphone distance and the intensity of the pop noise detected via microphone have inverse relationship with each other. The intensity of the recorded pop noise cannot be high if the distance between the microphone and the speaker is large enough (generally $> 50\text{ cm}$). This phenomenon can be used as an acoustic signature of the live speaker. Generally, the attacker may not be able to place the recording device near the speaker, which leads to the absence of the pop noise in the recorded voice. Hence, detection of *pop noise* can provide genuine acoustic cues for VLD, which will further be able to distinguish between the live (genuine) speech and replayed speech. Thus, VLD can be used to prevent the spoofing attacks.

The main task of the VLD system is to detect attributes of the live speaker, which is present in front of the ASV system. To that effect, the presence of the pop noise in the speech signal of a live speaker, is used as a discriminative acoustic feature. Thus, the speech is presented to the VLD system, where the pop noise detection algorithm detects the presence of the pop noise in the speech signal. If pop noise in the speech signal is detected, then the signal is passed to the ASV system for the verification. Figure 5.2 represents a schematic of ASV system being assisted by the VLD system.

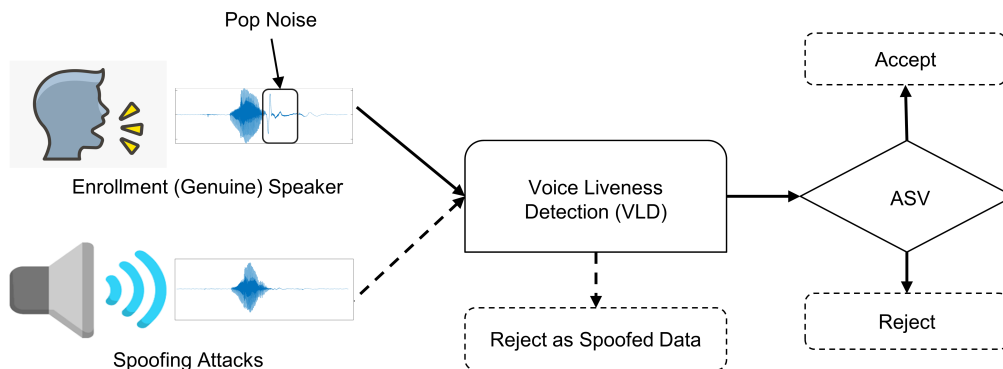


Figure 5.2: A Schematic of VLD System in Tandem with ASV System. After [34].

5.2.2.2 STFT-Based Baseline Algorithm

In [34], the features for pop noise detection are derived from the conventional STFT. The same algorithm is used in [8] for VLD using the POCO dataset. Hence, we consider it as a baseline approach, and consequently modified it to efficiently detect the liveness in the speech signal. In this baseline, energy spectral density (ESD) of the speech signal is estimated using spectrogram. Let $x(n)$ be the discrete-time input speech signal. Then, discrete-time STFT of $x(n)$ is calculated as [42]:

$$X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x(n) \cdot w(n, \tau) \cdot e^{-j\omega n}, \quad (5.1)$$

where $w(n, \tau)$ represents the analysis window, centered at time τ . It should be noted that $w(n, \tau)$ is a function of only time parameter, τ as independent variable. Eq. (5.1) can be rewritten as:

$$X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x(n, \tau) \cdot e^{-j\omega n}, \quad (5.2)$$

where $x(n, \tau) = x(n) \cdot w(n, \tau)$ is the windowed speech segment. Now, the spectrogram (i.e., ESD) is obtained by calculating the magnitude square of $X(\omega, \tau)$, i.e.,

$$S(\omega, \tau) = |X(\omega, \tau)|^2. \quad (5.3)$$

Here, $S(\omega, \tau) \in L^2(\mathbb{R}^2)$ (i.e., Hilbert space of finite energy signals over \mathbb{R}^2) [200]. Next, $S_{eng}(\omega, \tau)$ is calculated by considering ESD from $S(\omega, \tau)$ ranging within the frequency bins corresponding to $[0, \omega_{max}]$, i.e.,

$$S_{eng}(\omega, \tau) = S(\omega, \tau)_{0 \leq \omega \leq \omega_{max}}. \quad (5.4)$$

Here, ω_{max} is the digital frequency in *rad/s*, and let f_{max} be the corresponding frequency in *Hz*. Since the pop noise is observed in the lower frequency region of the spectrogram features, f_{max} may vary between 40 – 100 Hz. Let S_{avg} be the average of the ESD for each frame and computed as:

$$S_{avg}(\tau) = \frac{1}{N_b} \sum_{\omega=0}^{\omega_{max}} S_{eng}(\omega, \tau), \quad (5.5)$$

where N_b represents number of frequency bins corresponding to f_{max} Hz. Mean and standard deviation are estimated for averaged ESD $S_{avg}(\tau)$ in order to normalize it. As pop noise event lasts for a very short period of time (typically 20-100

ms) [147], 10 speech frames with the largest ESD were chosen using the $S_{avg}(\tau)$ and corresponding speech frames in $S_{eng}(\omega, \tau)$ are considered as a feature representation for pop noise detection². This feature set with appropriate labels is fed to a suitable classifier. The more details of this baseline algorithm can be found in [34].

²Empirically, it was observed that the performance of the proposed CQT-based feature set is fairly consistent *w.r.t.* variation in number of frames. In addition, the selection of 10 speech frames per utterance is also suitable choice for the fair comparison with STFT-based algorithm in [34], which also uses 10 speech frames per utterance. Hence, we considered 10 speech frames as an optimum value for our further experiments in this chapter.

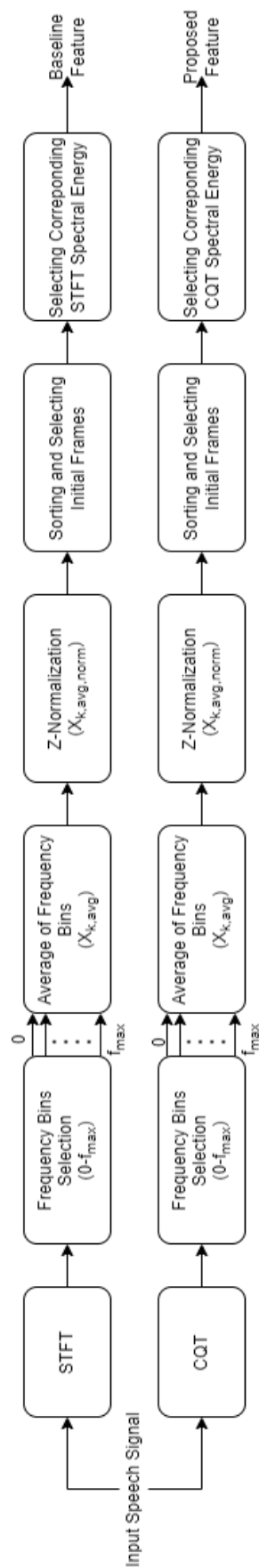


Figure 5.3: Functional Block Diagram of Baseline and Proposed Algorithm. After [35].

5.2.3 Proposed CQT-Based Algorithm

In the proposed approach, we employ CQT instead of STFT in order to obtain the high resolution frequency bins in the low frequency regions. The functional block diagram of the proposed algorithm along with baseline is shown in Figure 5.3. Following Brown's original approach [14], the frequency bins in CQT are geometrically-spaced as opposed to the linear spacing of bins in the STFT. By selecting the appropriate parameters of the CQT, we can locate the fine structural details of the spectrum of the pop noise, which are lying at very low frequency regions. Because of the geometrical spacing between the frequency bins in CQT, the low frequency region is well emphasized. For a time-domain signal, $x(n)$, CQT maps it into the time-frequency representation such that the quality factor, Q remains constant, and the center frequencies of the frequency bins are geometrically-spaced. Moreover, such constant Q analysis of the speech signals is desirable from both theoretical and practical viewpoints. In particular, CQT helps to preserve *form-invariance* property, such as the *linear* time-scaling property of the CTFT, which does not hold for STFT (because the analysis window used in STFT is a function of *only* the time parameter, τ as in eq. (5.1)). Furthermore, such form-invariance property is desirable for pattern recognition applications, where we want feature descriptors of a pattern to be *invariant w.r.t.* scale, shift, rotation, shape, etc. [200].

5.2.3.1 Development of CQT

Next, we develop expression for CQT. In the signal processing literature, discrete Fourier transform (DFT) is nothing but a uniformly sampled version of discrete-time Fourier transform (DTFT) [257]. In particular,

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x(n) \cdot e^{-j\omega n}, \quad (5.6)$$

and

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j(\frac{2\pi}{N})kn}, \quad (5.7)$$

where $X(e^{j\omega})$ and $X(k)$ represents DTFT and DFT of a discrete-time signal, $x(n)$, respectively. Thus, eq. (5.6) and eq. (5.7) are related by $\omega = (\frac{2\pi}{N}k) = \frac{2\pi}{T} = \frac{2\pi}{(N/k)}$, implying period in samples is $T = \frac{N}{k}$ and hence, number of cycles analyzed is equal to k . For speech signal processing applications, the speech signal is segmented using appropriate fixed window length at the segmental-level (i.e., fram-

ing) and appropriate % frame overlap. Each segment is then multiplied using a suitable window function (such as rectangular, Hamming, Hann, Gaussian, etc.) to avoid the *spectral leakage*. It should be noted that the role of the window function is to modify the *shape* of the segment of the speech signal. Let $x_s(n)$ and $w(n)$ represents the segment of the speech signal and window function, respectively, then DFT of the windowed signal is represented as:

$$X_s(k) = \sum_{n=0}^{N-1} x_s(n) \cdot w(n) \cdot e^{-j(\frac{2\pi}{N})kn}. \quad (5.8)$$

The index of each segment of the speech signal is supposed to have range from $n = 0$ to $n = N - 1$ for DFT computation. The DFT is performed on each segment of the speech signal $x(n)$, and it leads to STFT. It can be observed from eq. (5.7) that the frequency resolution or bandwidth (Δf) for DFT is equal to the sampling rate (F_s) divided by the window size (i.e., number of samples analyzed via time-domain window). Thus, in order to have the ratio of frequency (f) to bandwidth (Δf) to be constant (called as *constant Q*), the window size (δt) in the time-domain must vary *inversely* with frequency. In particular, we have,

$$\text{Quality factor } (Q) = \frac{\text{Center frequency } (f)}{\text{Bandwidth } (\Delta f)}. \quad (5.9)$$

For Q to be fixed, we have,

$$\Delta f \text{ (frequency resolution)} = \frac{\text{Sampling Rate } (F_s)}{\text{Window size } (\Delta t) \text{ in time - domain}}. \quad (5.10)$$

Since the sampling rate (F_s) of given data is fixed, eq. (5.10) can be written as:

$$\Delta t \cdot \Delta f \geq \text{constant}. \quad (5.11)$$

Eq. (5.11) is a manifestation of Heisenberg's uncertainty principle in signal processing framework (details given in Appendix A, where Δt and Δf are represented in the form of temporal variance σ_t^2 , and frequency variance σ_ω^2 , respectively). Let us consider Δt as $N(k)$, i.e., length of time-domain function in samples at frequency, f_k . From eq. (5.10), we have,

$$\Delta f_k = \frac{F_s}{N(k)}, \quad (5.12)$$

$$\therefore N(k) = \frac{F_s}{\Delta f_k}, \quad (5.13)$$

$$\because Q = \frac{f_k}{\Delta f_k} \implies \Delta f_k = \frac{f_k}{Q}. \quad (5.14)$$

Using eq. (5.14) in eq. (5.13), we get,

$$N(k) = \left(\frac{F_s}{f_k}\right) \cdot Q = T_k \cdot Q, \quad (5.15)$$

where $T_k = \frac{F_s}{f_k}$ is the period in samples, and it can be observed from eq. (5.15) that window $N(k)$ contains Q number of complete cycles for the k^{th} frequency component, f_k . From eq. (5.15), we have,

$$\frac{N(k)}{Q} = T_k. \quad (5.16)$$

Comparing eq. (5.16) with frequency spacing in the traditional DFT (where, $\omega_{DFT} = \frac{2\pi k}{N}$), we get the frequency spacing for CQT (ω_{CQT}) as:

$$\omega_{CQT} = \frac{2\pi}{\frac{N(k)}{Q}} = \left(\frac{2\pi}{N(k)}\right) \cdot Q. \quad (5.17)$$

Using eq. (5.7) and eq. (5.17), we obtain,

$$X_s(k) = \sum_{n=0}^{N(k)-1} x_s(n) w_k(n) e^{-j\left(\frac{2\pi}{N(k)} Qn\right)}. \quad (5.18)$$

The window function $w_k(n)$ has the identical shape for analysis of each frequency component f_k , however, its length is determined by $N(k)$ and thus, it is a function of both time and frequency bin index. Furthermore, because number of terms in each $X_s(k)$ varies with k , we must normalize *sum* in eq. (5.18) and thus, we get the expression of CQT as [14]:

$$X_s^{CQT}(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} x_s(n) w_k(n) e^{-j\left(\frac{2\pi}{N(k)} Qn\right)}. \quad (5.19)$$

The earlier original investigations have shown that since the time-domain window $w_k(n)$ is a function of both time and frequency parameters, the resulting transform integral yields constant Q (or constant percentage bandwidth) analysis, and also obeys form-invariance property [253]. Q is the quality factor, which is the ratio of center frequency to the bandwidth of each window, and it is given by eq. (5.14) [14]:

$$\because Q = \frac{f_k}{\Delta f_k} \implies \Delta f_k = \frac{f_k}{Q}. \quad (5.20)$$

$$\therefore Q = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/B} - 1}, \quad (5.21)$$

where B represents the number of bins per octave, and f_k represents the frequency of k^{th} the spectral component, which is given by:

$$f_k = (2^{\frac{k-1}{B}})f_{min}, \quad (5.22)$$

where f_{min} is the minimum frequency of the signal. Here, f is varied from f_{min} to f_{max} , which is chosen to be below Nyquist rate. Furthermore, we have used the resampling method in [172] to convert the geometrically-spaced frequency scale to linearly-spaced in order to obtain lower-dimensional feature representation. It helps to avoid computational complexity. In particular, the original CQT requires 567 frequency bins to represent the 0 to 40 Hz frequency range, whereas the uniformly resampled CQT requires only 118 frequency bins.

5.2.3.2 Spectrographic Analysis

As explained in the sub-Section 5.2.2.1, pop noise is a low frequency signal. Hence, it's spectral details also lie in the low frequency regions. The STFT transforms the speech signal into time-frequency domain, which possesses the constant separation between the frequency bins because STFT has constant resolution in the entire time-frequency plane (as per Heisenberg's uncertainty principle in signal processing framework [200]). However, the CQT displays the frequency-domain representation with high frequency resolution at lower frequency regions and vice-versa. Hence, CQT efficiently captures the spectral details of the pop noise. This can be observed from the waterfall plot for word "laugh" and it's top view of the STFT- and CQT-gram for the genuine *vs.* spoof speech signal as shown in Figure 5.4. The parameters of the CQT are tuned in order to emphasize the lower frequency regions, where the spectral details of pop noise are lying. The rectangular box in Figure 5.4 and Figure 5.5 represents the intended portion of the pop noise, whereas the encircled area represents the fundamental frequency (F_0) of the speech signal and its harmonics. It can be observed that the CQT gives more emphasis on pop noise region than the F_0 and its harmonics, as compared to its STFT counterpart. It can be clearly observed from Figure 5.4 that the CQT-gram emphasizes the pop noise vividly as compared to the traditional STFT-gram. The higher resolution property of the CQT allows the pop noise to occupy more area in the CQT-gram with higher intensity as compared to the STFT-gram. Because of having the larger region for pop noise in CQT-gram, it is much easier (than

its STFT counterpart) for the back-end classifier to discriminate the live *vs.* spoof speech signal. From Figure 5.4, it can be observed that the difference between the genuine *vs.* spoof speech signal is much more vivid in CQT as compared to its STFT counterpart. The non-linear geometrical spacing between the frequency bins can be clearly observed for the CQT-gram as opposed to the linear separation in STFT-gram. Furthermore, we have illustrated similar plots for the word “chip” in Figure 5.5. It can be observed that the difference in the spectral regions is very little for both the feature sets, which gives a cue that the probability of presence of pop noise in word “chip” is lower and hence, we get a relatively poor performance for this utterance (which will be discussed in the sub-Section 5.2.5).

5.2.3.3 CQT *vs.* STFT for Pop Noise Detection

Table 5.1 represents the comparison of CQT, *resampled* CQT, and STFT for various parameters utilized in this study. As the pop noise is present in the lower frequency regions (in particular, up to 40 Hz), we show the number of bins required to represent the frequency range $f_{min} - f_{40Hz}$. It can be observed from Table 5.1 that 567 frequency bins are required for CQT to represent the frequency range of $f_{min} - f_{40Hz}$. However, *resampled* CQT takes 118 frequency bins to represent the frequency range of $f_{min} - f_{40Hz}$. In our previous work [35], we utilized the 1 frequency bin per Hz for the STFT to represent the Nyquist frequency range, which is shown as STFT-1 in Table 5.1. For fair comparison, we also performed the experiments for STFT with 3 frequency bins per Hz as the *resampled* CQT utilizes the similar frequency resolution. In Table 5.1, this high frequency resolution STFT is denoted as STFT-2. However, empirically, it has been observed that the STFT-based features with 1 frequency bins per Hz gives the better performance. Hence, the results reported in this thesis for STFT utilizes the frequency resolution of 1 frequency bin per Hz.

The MATLAB pseudocode of the proposed CQT-based approach is shown in Algorithm 5. Here, f_{CQT} is obtained by uniform resampling of log-power magnitude CQT spectrum. ESD S_{eng} is obtained by considering frequency bins (f_{bins}) within the interval $[0, f_{max}]$ and taking absolute of it. As pop noise is present at low frequency region, we have varied f_{max} from 10 Hz to 100 Hz in order to observe the effect of presence of pop noise. Furthermore, the average of CQT spectrogram $f_{k,avg}$ is computed for each column vector in order to obtain it frame-wise. Furthermore, we found the Q -factor pertaining to pop noise detection is $Q = 134$ as opposed to $Q = 34$ for analysis of western music as in Brown’s original work [14]. Then normalization of $f_{k,avg}$ is done to zero-mean and unit stan-

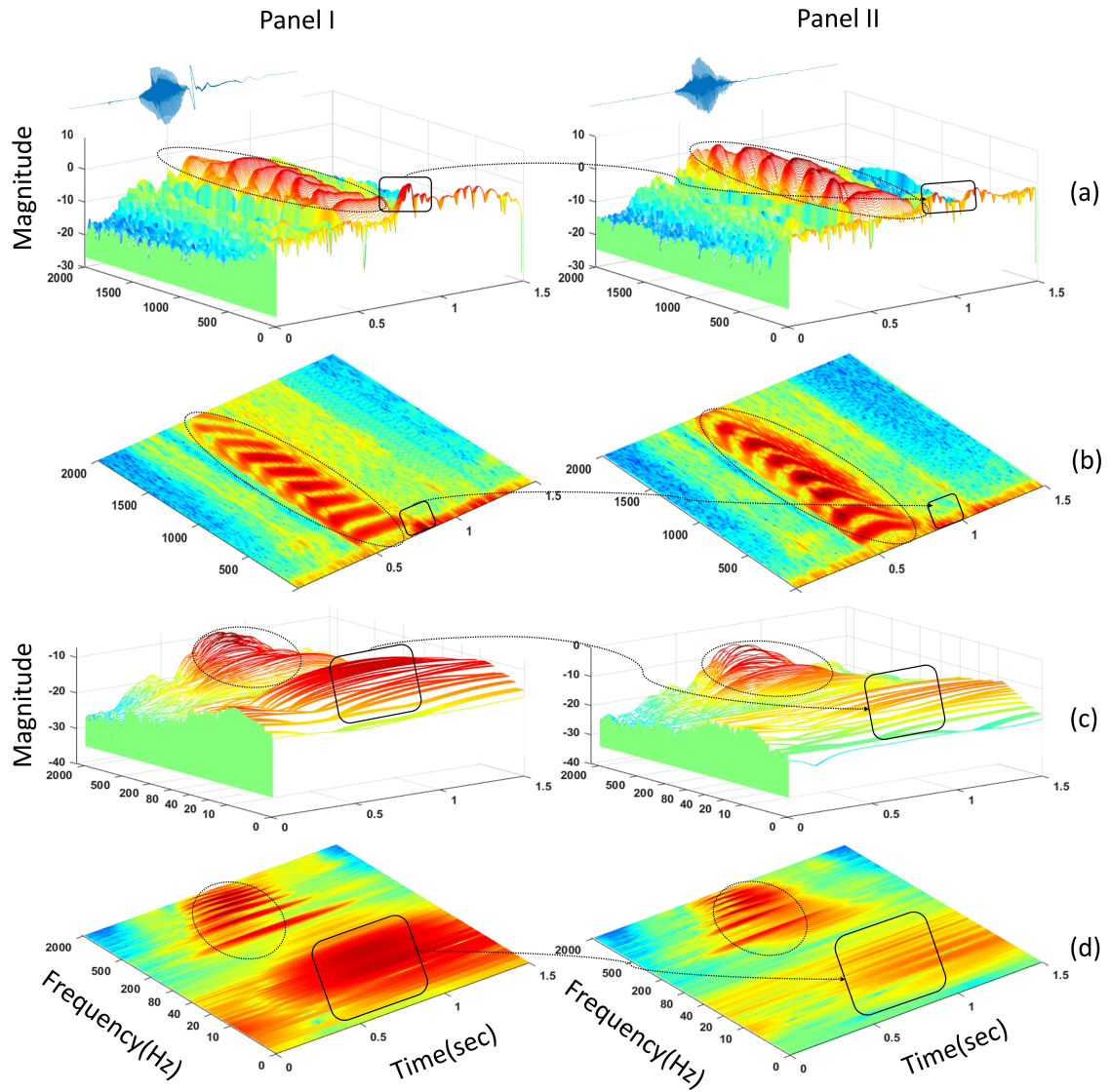


Figure 5.4: Panel-I and Panel-II Depicts the Spectrographic Analysis for Genuine *vs.* Spoof Speech Signal for Word "Laugh", Respectively. (a) the Waterfall Plot for STFT, (b) the Top-view of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the Top-view of the CQT Waterfall Plot. The Rectangular Box Represents the Intended Location of the Pop Noise, whereas the Encircled Region Represents the Presence of F_0 and Its Harmonics. After [15].

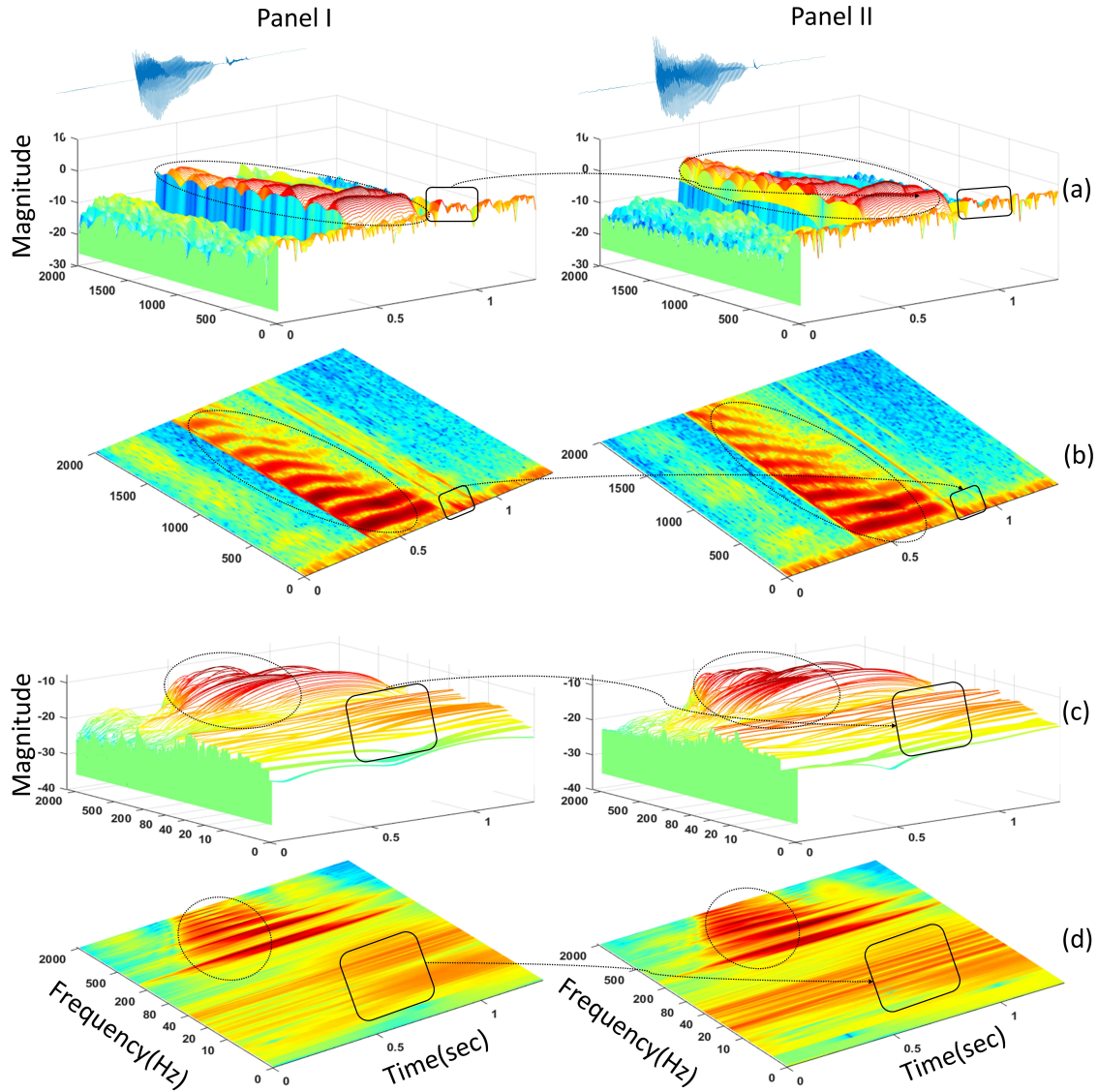


Figure 5.5: Panel-I and Panel-II Depicts the Spectrographic Analysis for Genuine *vs.* Spoof Speech Signal for Word "Chip", Respectively. (a) the Waterfall Plot for STFT, (b) the Top-view of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the Top-view of the CQT Waterfall Plot. The Rectangular Box Represents the Intended Location of the Pop Noise, whereas the Encircled Region Represents the Presence of F_0 and Its Harmonics. After [15].

Table 5.1: The Comparison of the CQT, *Resampled CQT*, and STFT *w.r.t.* the Various Spectrographic Parameters for Pop Noise Detection with $F_s = 22050$ Hz. After [14].

Parameter	CQT [14]	Uniformly Resampled CQT [172]	STFT-1 [35]	STFT-2 [15]
f_{min}	0.67 Hz	0.67 Hz	1 Hz	0.33 Hz
$f_{Nyquist} = \frac{F_s}{2}$	11050 Hz	11050 Hz	11050 Hz	11050 Hz
f_{40Hz}	40 Hz	40 Hz	40 Hz	40 Hz
Number of Frequency Bins for $f_{Nyquist}$	1345	32850	11050	33075
Number of Frequency Bins for f_{40Hz}	567	118	40	120
Resolution	varying = $\frac{f \cdot k}{Q}$	0.3365 Hz	1 Hz	0.3333 Hz
Quality Factor (Q)	Constant	Variable	Variable	Variable

dard deviation to obtain $f_{k,avg,norm}$. As pop noise lasts for a very short period of time [34], 10 frames (equivalent to 20 ms) from $f_{k,avg,norm}$ having the largest ESD is considered, and then taking frames corresponding to that indices from S_{eng} in order to obtain the ESD of the pop noise region.

In the anti-spoofing literature, the cepstral features, namely, CQCC and LFCC have shown good performance for SSD. Hence, in this thesis, we extended the experiments on POCO dataset. CQCC is obtained by performing the DCT operation on the uniformly resampled CQT [172]. Resampling is necessary to linearize the scale of the CQT so that DCT can be applied [258]. In CQT, frequency bins are geometrically-spaced, whereas the frequencies of the basis functions in the DCT are linearly-spaced. Resampling is performed on CQT to make the scale of the CQT similar to that of the frequencies of the basis functions in the DCT. Resampling allows us to extract the cepstral coefficients of CQT by preserving the orthogonality condition of the DCT basis functions. 90-dimensional (i.e., 90-D) CQCC consists of static, Δ , and $\Delta\Delta$ features and it is extracted using the desirable parameters for this application, i.e., $f_{min} = 0.67$ Hz, and $B = 96$. The LFCC feature set is extracted by applying the triangular-shaped linearly-spaced subband filters on STFT spectrogram and then followed by DCT operation. 40 subband filters are utilized in the filterbank and all 40 cepstral coefficients are retained and appended with Δ and $\Delta\Delta$ coefficients to form 120-D LFCC feature set.

Furthermore, we exploit Heisenberg's uncertainty principle in signal processing framework to analyze the temporal variance (σ_t^2) and frequency variance (σ_ω^2)

Algorithm 5 MATLAB Pseudo Code of Proposed CQT-based Algorithm. After [15].

- | | |
|---|--|
| <ol style="list-style-type: none"> 1. $f_k = (2^{\frac{k-1}{B}})f_{min}$, 2. $N(k) = \frac{F_s}{\Delta f_k}$, 3. $X_s^{CQT}(k) = \langle x_s(n) \cdot w(n), e^{j\frac{2\pi Qn}{N(k)}} \rangle$, 4. for $i = 1 : N_{columns}(X_{CQT})$ do,
 $X_{CQT}(k, i) = X_{x_{si}}^{CQT}(k)$, <li style="padding-left: 2em;">end for 5. $S_{eng} = (abs(X_{CQT}(1 : X_{bins}(f_{max}), :)))^2$, 6. for $i = 1 : N_{frames}(S_{eng})$ do,
 $X_{k,avg}(i) = mean(S_{eng}(:, i))$, <li style="padding-left: 2em;">end for 7. $MN = mean(X_{k,avg})$, $SD = std(X_{k,avg})$, 8. for $i = 1 : N_{frames}(S_{eng})$ do
 $X_{k,avg,norm}(i) = (X_{k,avg}(i) - MN) / SD$, <li style="padding-left: 2em;">end for 9. $[X_{k,avg,norm,sort}, index] = sort(X_{k,avg,norm})$, 10. $X_{k,avg,initial} = X_{k,avg,norm,sort}(1 : 10, :)$, 11. $index_{initial} = index(1 : 10, :)$, 12. $CQT_{features} = S_{eng}(:, index_{initial}(i))$, | <p>geometrically-spaced frequency bins,</p> <p>computation of CQT,
 framewise concatenation of CQT,
 CQT computed for corresponding
 segment x_{si} for i^{th} column,</p> <p>Taking bins up to f_{max} only,
 N_{frames} corresponds to
 number of frames,
 Taking average of CQT spectrogram
 along frequency bins,</p> <p>Estimate mean and standard
 deviation,</p> <p>Normalizing,</p> <p>Sorting,
 Taking initial 10 frames,
 Taking corresponding indices,
 Feature set</p> |
|---|--|
-

of the analysis window function for CQT and STFT. As per the uncertainty principle, $\sigma_t^2 \cdot \sigma_\omega^2 \geq \frac{1}{4}$ (proof is given in the Appendix A). The area $\sigma_t^2 \cdot \sigma_\omega^2$ is called as Heisenberg's box (or TBP) in the time-frequency plane [200]. In the traditional STFT, the area of Heisenberg's box for the analysis window always remains constant (in the entire time-frequency plane) as opposed to the CQT, where the lower frequency region possesses higher frequency resolution and lower temporal resolution and vice-versa [202]. The Table 5.2 shows the length of the analysis window for the various analysis frequencies in the CQT *w.r.t.* the CQT parameters utilized in this study. It can be clearly observed from Table 5.2 that the length of the analysis window is very large for lower frequencies, and vice-versa. However, in STFT, the length of the analysis window remains constant across the entire time-frequency plane [200].

Furthermore, we analyzed the TBP for CQT at various frequencies using several analysis windows, namely, Hamming, hann, and Gaussian. The hann window is also known as *raised cosine* [259]. The TBP for the analysis window is

Table 5.2: Window Length in Samples as a Function of Analysis Frequency (f_k). After [14].

k	Frequency (Hz)	Window (Samples)	Duration (s)
1	0.6729	4390912	199.13
10	0.7181	4114652	186.60
100	1.3753	2148411	97.43
200	2.83	1043625	47.32
300	5.82	506957	23
400	12	246262	11.16
500	24.69	119626	5.42
600	50.84	58110	2.63
700	104	28228	1.28
800	215	13712	0.62
1200	3870	763	0.0346
1300	7970	370	0.0168
1345	11025	268	0.0122

computed as the product of the temporal and frequency variances of the analysis window. Figure 5.6(a), Figure 5.6(b), and Figure 5.6(c) shows the temporal variance, frequency variance, and TBP for various window functions. It can be observed that the TBP for hann and Gaussian window is constant *w.r.t.* analysis frequency, whereas it goes on decreasing for Hamming.

5.2.4 Experimental Setup

5.2.4.1 Dataset Used

In this work, we have used the POCO dataset. The detailed discussion of the POCO dataset can be studied in Chapter 3 (Section 3.2.6). We partitioned the speech samples from RC-A and RP-A into three subsets, namely, training, Dev, and Eval. The dataset is partitioned with a ratio of 40 %, 20 %, and 40 % into these subsets. We also ensured that the speakers are exclusive in each subset and the ratio between male and female speakers is maintained. The detailed statistics of this data distribution is given in Table 5.3. However, an adapted version of the dataset by including the simulated replay mechanism is explained in the following Section.

To investigate the effect of replay spoof mechanism in the context of pop noise, we have generated the simulated replay dataset from the POCO database. The procedure followed for this is the same as that followed for the generation of the

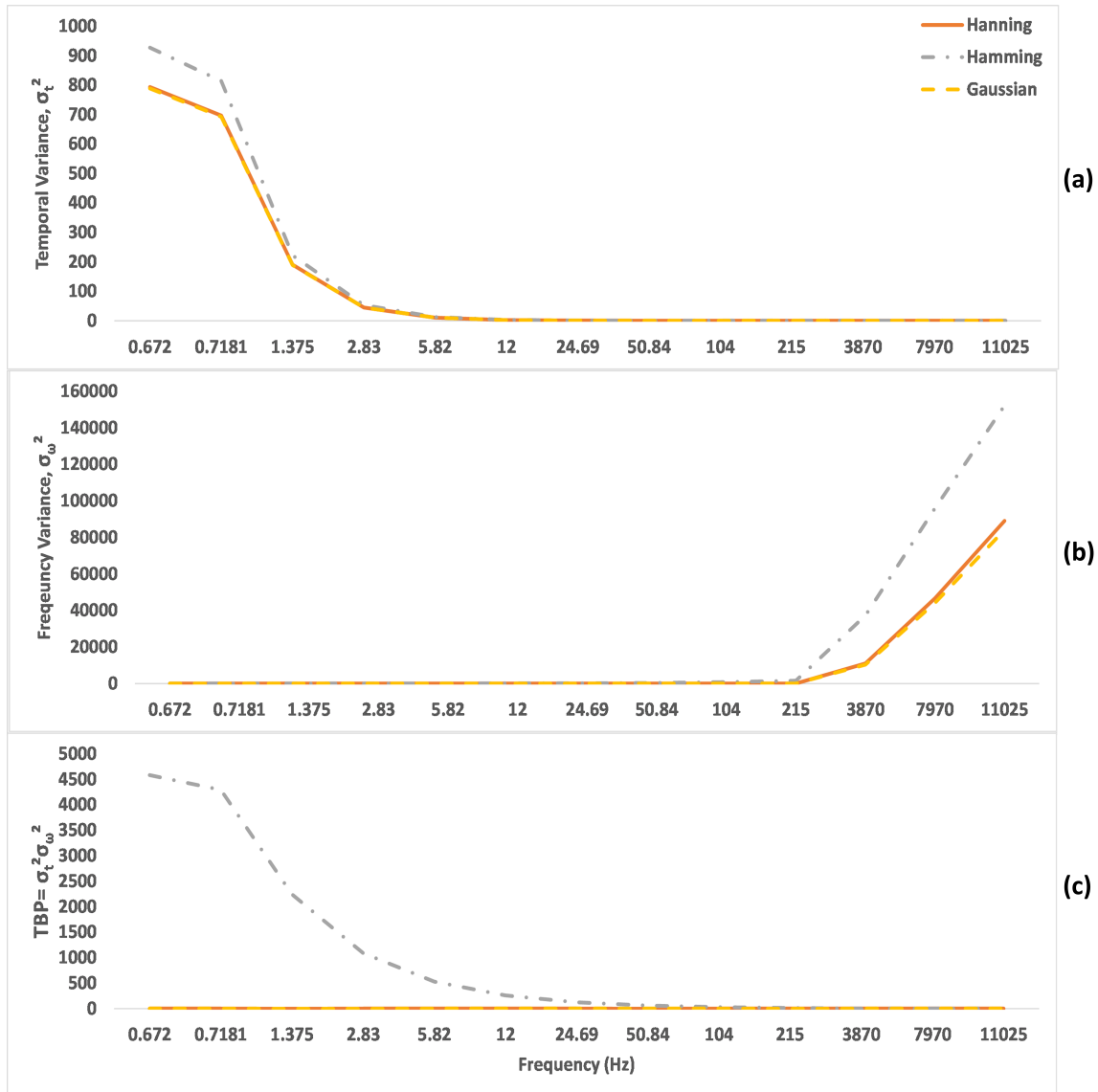


Figure 5.6: (a) Temporal Variance (σ_t^2), (b) Frequency Variance (σ_ω^2), and (c) TBP (i.e., $\sigma_t^2 \cdot \sigma_\omega^2$) for Hamming, Gaussian, and hann Windows. After [15].

Table 5.3: Statistics of the POCO Dataset used for Experiments in this Thesis. After [8, 15].

Subset	# Utterances	# Speaker	# Male	# Female
Training	6952	27	13	14
Dev	3432	13	6	7
Eval	6600	26	13	13

replay speech samples in ASVSpooof 2019 challenge dataset [5]. The details can be studied in [260, 261]. The replayed speech is generally known to observe the reverberation effect of the associated acoustic medium. To replicate this effect, the geometrical acoustics is generated by using an image-source model equivalent to a perfect rectangular parallelepiped room to churn out an impulse response to a directional receiver from each omni-directional primary source [36]. It is assumed that the replayed speech is first recorded by a microphone before it is being replayed from the non-linear replay device. Figure 5.7 illustrates the schematic diagram for generating synthetic replay using image-source model, in particular, the manner in which the images of the source are spatially arranged. The highlighted rectangle represents the original room.

For the generation of simulated replayed speech, it is required to define the surface material of the room, it's dimensions, the location of the primary source, and the receiver system. Then, the impulse response of the simulated room is estimated using the reverberation time (RT), which is calculated using Norris-Eyring formula given by [16, 262, 263]:

$$RT = \frac{KV}{-N \ln(1 - \alpha_m)}, \quad (5.23)$$

where K is a constant determined by Sabine's formula [264], and is taken as 0.161, V is the volume of the room, N is a number of surfaces in the room, and α_m is the average coefficient of absorption, which is defined as:

$$\alpha_m = \frac{\sum_{n=1}^{n=N} s_n \alpha_n}{N}, \quad (5.24)$$

where s_n and α_n is the n^{th} element of surface, and the corresponding coefficient of absorption, and $N = s_1 + s_2 + \dots + s_N$.

Then, the image source-to-receiver responses are calculated using the method of image source [36]. The individual image source responses are gathered to obtain the complete impulse response from each primary source to the receiver. Fi-

Table 5.4: Parameter and Corresponding Configuration for Replay Mechanism. After [16].

Parameter	Configuration
Room Size	$3.55 m^2$
Sensor Position	(2,1,1.4)
Source Directivity	Omnidirectional
Sensor Directivity	Omnidirectional
Reverberation Time	0.07 sec

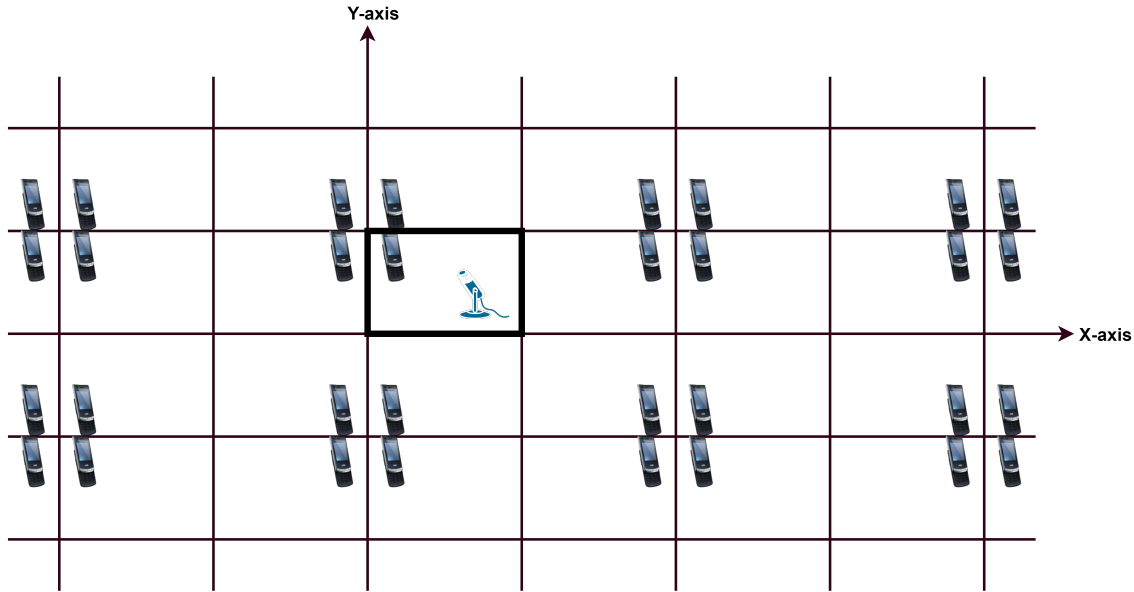


Figure 5.7: Schematic Diagram for Generation of Synthetic Replay using Image Source Model. After [36].

nally, the obtained two-channel impulse response is convolved with the audio files to obtain the synthetic replayed speech, which will approximate to originally replayed speech. Table 5.4 represents the parameter and corresponding configuration assumed for the replay mechanism. The source position is varied from $10 cm$ to $90 cm$ for observing the effect of variation in distance between speaker and attacker's recording device. We have added the replay mechanism on RP-A subset to obtain the new replayed subset, named as, *REP-A*.

The experiments using proposed CQT-based feature sets are extended on ASVSpooof challenge datasets, namely, ASVSpooof 2019 PA and ASVSpooof 2017 version 2.0 dataset. These are popular datasets in the literature for replay anti-spoofing. The detailed statistics of the ASVSpooof 2019 PA dataset and ASVSpooof 2017 version-2 dataset are shown in Table 3.5 and Table 3.3, respectively. Further details of these datasets can be studied in [3,4].

5.2.4.2 Classifiers Used

In this study, we utilized the GMM, SVM, CNN, LCNN, and ResNet as classifiers. The CNN used in this work consists of four convolution layers, and 1 FC layer. The output of these four convolutional layers have 4, 16, 32, and 8 channels, respectively. The convolution operations are done using kernel size of 3x3. In addition, convolution operation is performed using zero-padding with a stride of 1. The final convolution block is followed by a fully-connected linear layer with 1312 hidden units. The output of the final layer is activated using a sigmoid function, which makes the final decision of whether the utterance contains pop noise or not. ReLU function is used as the activation function in the hidden layers. The model is trained using Stochastic Gradient Descent (SGD) algorithm with a batch size of 64, and learning rate of 0.001. Binary cross-entropy loss is chosen as the loss function. The experiments are executed for a total number of 400 epochs. LCNN architecture uses four convolutional layers, each followed by MFM activation function. The fully-connected FC5 layer contains a low-dimensional high-level audio representation. Then, the FC6 layer with softmax activation function was used to distinguish between spoofing and genuine classes during the training process. The details of LCNN architecture utilized for CQT-based feature set are shown in Table 5.5. ResNet is employed for this task to take the advantage of high-level features.

5.2.5 Experimental Results

In this Section, we evaluate the performance of CQT-based proposed VLD system for different evaluation factors. In addition, we have also compared the proposed approach with cepstral-based features, namely, LFCC and CQCC. Furthermore, we have analyzed the effect of variation in the frequency range, wordwise classification accuracy, and phoneme-based performance. We have also shown the analysis by considering REP-A subset in which we have embedded the simulated replay mechanism in order to observe the performance for a realistic anti-spoofing scenarios. The results are reported using % Classification Accuracy and % EER.

5.2.5.1 Effect of Variation in Frequency Range

In this Section, we illustrate the effect of variation in the frequency range for lower frequencies (as pop noise is present in the low frequency region). The experiment is performed using RC-A and RP-A subsets, explained in Chapter 3, Section 3.2.6. The system is designed using the feature set derived from the CQT along with

Table 5.5: Details of the Proposed LCNN Architecture for VLD. After [17].

Layer	Filter/Stride	Output	# Parameters
Conv1	5x5/1x1	8 x 88 x 8	205
MFM1	-	4 x 88 x 8	-
Conv2a	1x1/1x1	8 x 88 x 8	40
MFM2a	-	4 x 88 x 8	-
Conv2b	3x3/1x1	32 x 86 x 6	1184
MFM2b	-	16 x 86 x 6	-
Conv3a	1x1/1x1	32 x 86 x 6	544
MFM3a	-	16 x 86 x 6	-
Conv3b	3x3/1x1	64 x 84 x 4	9280
MFM3b	-	32 x 84 x 4	-
Conv4a	1x1/1x1	64 x 84 x 4	2112
MFM4a	-	32 x 84 x 4	-
Conv4b	3x3/1x1	16 x 82 x 2	4624
MFM4b	-	8 x 82 x 2	-
FC5	-	1 x 128	168k
MFM5	-	1 x 64	-
FC6	-	1 x 1	65

SVM classifier. The details of the proposed feature set and classifier, are discussed in Section 5.2.3 and Chapter 3 (Section 3.4.2), respectively. The experiments are performed with the variation of the lower range of frequencies, and the corresponding results are displayed in Table 5.7. It can be observed from Table 5.7 that the % classification accuracy for lower frequency range (i.e., 1-10 Hz, 1-20 Hz, and 1-30 Hz) is almost equal and is relatively maximum. As we increase the frequency, the performance of the pop noise detection system degrades. This finding validate the fact that the pop noise is predominantly present at low frequency regions, i.e., below 30 Hz for CQT-based feature set. Hence, we set the value of f_{max} to 30 Hz for the CQT-based features used in the further set of experiments.

Furthermore, experiments are performed for CQT *vs.* resampled CQT using SVM classifier by setting f_{max} to 30 Hz. It can be observed from Table 5.8 that the resampled CQT performs comparatively better than the original CQT. Considering these two advantages concerned with the VLD task (i.e., better performance with lower-dimensional feature set), we employed the resampled CQT-based feature representation for all the experiments in this chapter.

Table 5.6: Details of the Proposed ResNet Architecture for VLD. After [18].

Layer	Filter	Output
Conv1	$7 \times 7, 16$	90×10
Conv2	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 2$	90×10
Conv3	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$	45×5
Conv4	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	23×3
Conv5	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	12×2
FC6	-	768

Table 5.7: Results (in % Classification Accuracy) for CQT-SVM-based Pop Noise Detection using RC-A (genuine) *vs.* RP-A (spoof) Dataset with Variation in Frequency Range. After [15].

Frequency Range	Dev	Eval
1-10 Hz	79.77	78.39
1-20 Hz	79.95	78.26
1-30 Hz	79.34	78.42
1-40 Hz	77.88	74.92
1-50 Hz	68.04	64.80
1-60 Hz	79.17	74.47
1-70 Hz	71.15	67.86
1-80 Hz	78.55	74.71

Table 5.8: Results in (% Classification Accuracy) for the Original CQT-based Algorithm *vs.* Resampled CQT-based Algorithm using SVM Classifier on POCO Dataset. After [15].

CQT Version	Dev	Eval
Original CQT	74.98	74.50
Resampled CQT	79.84	78.88

5.2.5.2 Effect of Number of Frames

In this sub-Section, we analyzed the performance of the proposed feature set by varying the number of frames of the speech signal. It can be observed from Table 5.9 that, as we increase the number of frames, the classification accuracy increases as well. However, the improvement in classification accuracy by considering higher number of frames is not that significant when compared to the less

number of frames. Furthermore, it is interesting that the given feature representation produce the comparable performance using only a single frame. To have a lower dimension feature representation and due to the fact that pop noise lasts only for a short period of time (i.e., 20 - 100 ms [34]), we considered 10 frames as an optimum value for our further experiments. Furthermore, the selection of 10 frames per utterance is also suitable choice for the fair comparison with STFT-based algorithm, which also uses 10 frames per utterance.

Table 5.9: Results (in % Classification Accuracy) for Varying the Number of Frames in Proposed CQT-based Algorithm with SVM Classifier on POCO Dataset.

# Frames	Dev	Eval
1	79.02	78.30
2	79.02	78.38
3	79.08	78.45
4	79.11	78.38
5	78.9	78.45
10	79.34	78.42
20	79.81	77.83
30	79.28	78.22
40	79.78	78.74
50	79.53	78.86
60	79.84	78.76
70	80.16	79.09
80	80.3	79.18
90	80.71	79.48
100	80.65	79.8

5.2.5.3 Effect of Various Analysis Window Functions in CQT

In our recent work, we reported the experimental results with hann window as an analysis window in CQT for VLD task [151]. Experiments are also performed with the other window functions, namely, Hamming, and Gaussian in the proposed CQT-based feature set along with various classifiers, and results are reported in Table 5.10. It can be observed that the Gaussian window is more suitable for all the classifiers (considered in this study) as it shows the relatively better results for all the classifiers, on Eval set. This is in agreement with the fact that Gaussian window possess lowest TBP value amongst all window functions in the framework of Heisenberg’s uncertainty principle. As best possible results are obtained with

Table 5.10: Results (in % Classification Accuracy) of Proposed CQT-based Approach with Different Window Functions using Various Classifiers.

Feature Set	Classifier	Hann		Hamming		Gaussian	
		Dev	Eval	Dev	Eval	Dev	Eval
CQT	SVM	79.34	78.42	79.05	78.14	79.84	78.88
CQT	GMM	73.48	72.59	72.73	72.18	74.07	72.64
CQT	CNN	82.27	79.77	82.51	81.78	81.52	81.82
CQT	LCNN	83.68	81.93	83.91	81.89	84.84	82.45
CQT	ResNet	82.45	79.43	82.54	78.86	83.04	80.42

Gaussian window, further experiments are performed with the Gaussian window in CQT-based feature set.

5.2.5.4 Comparison of Results for STFT *vs.* CQT using Various Classifiers

In this sub-Section, we show the comparison of the results for proposed CQT-based algorithm with STFT-based baseline algorithm. For STFT-based baseline algorithm (SVM as a classifier), we utilized the similar approach as explained in [34], and the same approach was utilized in the original POCO dataset paper [8]. We could reproduce the results given in [8] on STFT-based feature set using cross-validation, and were reported in our earlier work [35]. Furthermore, we partitioned the dataset in our earlier work [35], and the similar subsets are utilized in this work. For the proposed CQT-based algorithm, f_{max} is tuned to 30 Hz, and analysis window is set to be Gaussian, as discussed in sub-Section 5.2.5.1 and sub-Section 5.2.5.3, respectively. From Table 5.11, it can be observed that when SVM is used as a classifier, the evaluation accuracy is 67.93 % for STFT-based baseline algorithm, whereas it is 78.88 % for the proposed CQT-based algorithm. Thus, there is an approximate 11.25 % of improvement from the baseline system. Also with GMM as a classifier, the % classification accuracy of 53.85 % is obtained on Eval set for the STFT-based baseline algorithm and % classification accuracy of 72.64 % is obtained using proposed CQT-based algorithm. Here, we obtain an absolute improvement of 18.79 % from the baseline algorithm. Furthermore, when CNN is used as a classifier, the baseline algorithm gives a classification accuracy of 71.81 %, which is 81.82 % for the proposed algorithm. Here also, the absolute improvement of 10.01 % in classification accuracy is obtained for the proposed CQT-based algorithm when compared with the STFT-based baseline algorithm. Furthermore, LCNN and ResNet classifiers shows the similar trends in performance for the baseline *vs.* proposed algorithm, and in particular, LCNN

Table 5.11: Comparison of Proposed CQT-based Approach with the STFT-based Baseline Approach using Various Classifiers. After [15].

Feature Set	Classifier	% Accuracy		% EER	
		Dev	Eval	Dev	Eval
STFT	SVM	65.61	67.93	37.61	35.11
STFT	GMM	55.22	53.85	40.42	41.60
STFT	CNN	70.57	71.81	31.80	29.15
STFT	LCNN	70.60	71.90	30.37	28.69
STFT	ResNet	72.05	71.84	34.34	33.86
CQT	SVM	79.84	78.88	20.49	21.37
CQT	GMM	74.07	72.64	25.70	26.52
CQT	CNN	81.52	81.82	18.67	18.25
CQT	LCNN	84.84	82.45	15.84	17.78
CQT	ResNet	83.04	80.42	23.64	22.96

shows the absolute improvement in classification accuracy of 0.77 % and 2.3 % for CQT-based algorithm over the CNN and ResNet classifiers, respectively. Overall, the proposed algorithm shows the significant improvement over the baseline algorithm for all the four classifiers (considered in this study) indicating utility of proposed feature set across various statistical (GMM), discriminative (SVM), and deep learning-based (CNN, LCNN, and ResNet) classifiers. In baseline algorithm, the maximum frequency (f_{max}) for pop noise detection was considered to be 40 Hz as it was given in [8, 34]. The performance of these systems is evaluated using the other evaluation metric, i.e., % EER. From Table 5.11, it can be observed that the similar trends in the results are observed for % EER.

Furthermore, we have performed the experiments on POCO dataset using cepstral-based feature sets, namely, CQCC and LFCC; since these feature sets are utilized as baseline systems for ASVSpooof challenges (i.e., ASVSpooof-2015, -2017, -2019, and -2021 challenge campaigns). The experiments are performed using all classifiers in this study, namely, GMM, SVM, CNN, LCNN, and ResNet and the results are reported in Table 5.12. It can be observed from Table 5.12 and Table 5.11 that the cepstral-based features (i.e., LFCC and CQCC) could not perform well for VLD task as compared to the STFT-based and proposed CQT-based feature sets.

Figure 5.8 shows the DET curves for the STFT-based baseline algorithm and CQT-based proposed algorithm along with GMM, SVM, CNN, LCNN, and ResNet as classifiers. Figure 5.8(a) and Figure 5.8(b) depicts the performance of the mentioned algorithms on Dev and Eval sets, respectively. It can be observed from Figure 5.8 that the DET curves shows the better performance for CQT-based al-

Table 5.12: Comparison of CQCC and LFCC Feature Sets using Various Classifiers on POCO Dataset. After [15].

Feature Set	Classifier	Accuracy		EER	
		Dev	Eval	Dev	Eval
CQCC	GMM	78.47	67.95	21.70	31.00
CQCC	SVM	63.69	57.67	35.75	41.27
CQCC	CNN	67.04	60.06	47.59	46.47
CQCC	LCNN	69.25	63.00	43.09	43.65
CQCC	ResNet	65.99	60.19	45.05	48.72
LFCC	GMM	81.82	72.47	18.23	26.59
LFCC	SVM	50	50	50	50
LFCC	CNN	68.91	61.87	31.07	37.66
LFCC	LCNN	66.84	60.37	33.39	39.08
LFCC	ResNet	67.33	60.59	32.67	39.05

gorithms over STFT-based baseline algorithm for all the five classifiers. Furthermore, proposed CQT-based feature set with LCNN gave relatively best performance. Moreover, these results using DET curves are in agreement with % classification accuracy for the same experiment as shown in Table 5.11.

Furthermore, the experiments were performed by increasing the amount of training data in order to majorly investigate the improvement in the performance for CNN, LCNN, and ResNet architectures, since deep learning architectures are known to perform well for a large amount of training data. To that effect, we divided the dataset into two parts, i.e., training and testing with a ratio of 80 % and 20 %, respectively. The corresponding results are presented in Table 5.13. It can be observed that the performance of both the baseline and proposed approaches with SVM as a classifier is almost the same as stated in Table 5.11, where dataset division is 40 % training, 20 % Dev, and 40 % Eval. However, the proposed CQT-based algorithm shows the marginal improvement over all the algorithms.

Next, the experiments are performed to obtain the wordwise % classification accuracy to observe the significance of presence of pop noise for each word. Figure 5.9 represents the comparison of wordwise accuracy on Dev set for baseline and proposed algorithm for SVM, GMM, CNN, LCNN, and ResNet. Similarly, Figure 5.10 represents the comparison of wordwise accuracy on Eval set for baseline and proposed algorithm. Here, for Eval set, it can be observed that for words, such as 'busy', 'division', 'fat', 'funny', 'five', 'thong', and 'shout', the accuracy is around 80 %, 55 %, 85 %, 90 %, and 85 % for the proposed CQT-based algorithm for SVM, GMM, CNN, LCNN, and ResNet, respectively. For the other words,

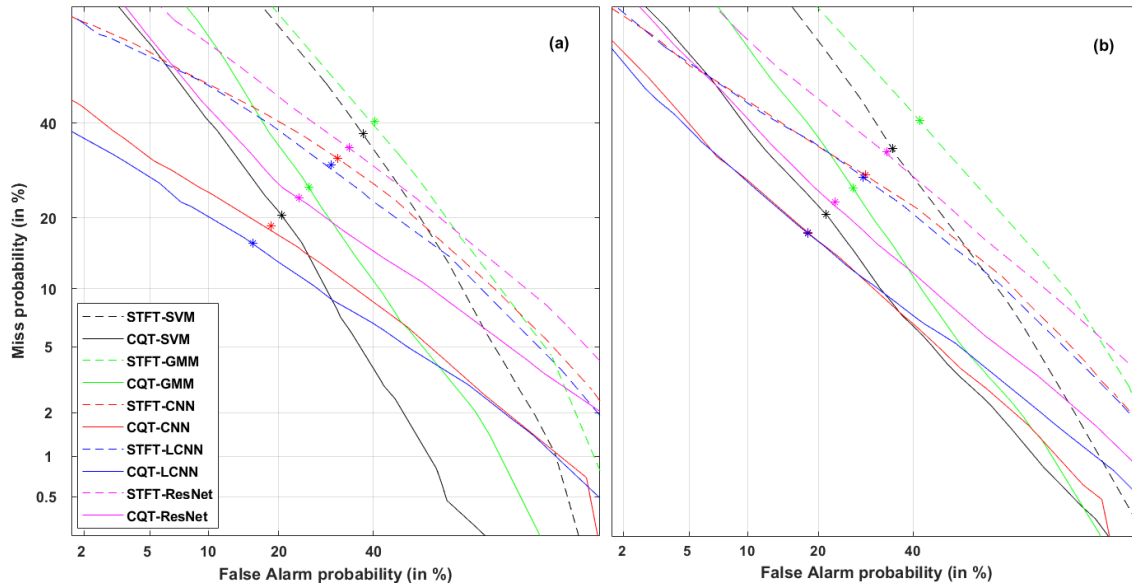


Figure 5.8: DET Curves for the Proposed CQT-based Algorithm *vs.* STFT-based Baseline Algorithm for Various Classifiers on (a) Dev, and (b) Eval Set. Legends in Figure 5.8(b) are the Same as that of Figure 5.8(a). After [15].

the accuracy is a bit lower, still when compared with the baseline algorithm, the proposed methods perform well except for the few words, such as ‘laugh’, ‘who’, and ‘wolf’. Furthermore, the % classification accuracy is computed *w.r.t.* various classes of phonemes, namely, affricate, fricative, plosive, and nasal, as shown in Table 5.14. The results are shown *w.r.t.* initial dataset division, i.e., 40 % training, 20 % Dev, and 40 % Eval.

It can be observed from Table 5.14 that average accuracy for affricate sound is relatively highest followed by nasal, fricative, and plosive sounds (where pop noise can be present at the start or can overlap with the sound, or it can occur at the end of the sound). Relatively better results for affricate sound can be justified by the fact that affricates are the counterpart of diphthongs and thus, they are transitional speech sounds, consisting of consonant-plosive-fricative combinations, rapidly transiting from plosives to fricatives. Moreover, production of plosive requires a sequence of events, in particular, complete closure of the oral tract, generation of turbulence for a very short-time duration (i.e., burst), generation of aspiration, and the onset of the following vowel. On the other hand, production of fricative involves generation of frication noise by turbulent airflow at some point of constriction created by the tongue, a constriction that is narrower than with the vowels [42]. These production characteristics of plosives and fricatives makes their spectrum to occupy high frequency region and hence, requires high temporal resolution of analysis window function. To that effect, it can be

Table 5.13: Comparison of Proposed Approach *vs.* the Baseline Approach with Larger Training Data (80 % Training, 20 % Testing) for Various Classifiers on POCO Dataset. After [15].

Feature Set	Classifier	Accuracy (%)	EER (%)
STFT	SVM	66.78	37.04
STFT	GMM	55.48	40.30
STFT	CNN	70.62	30.30
STFT	LCNN	71.76	28.85
STFT	ResNet	74.21	34.01
CQT	SVM	80.59	20.13
CQT	GMM	74.21	25.37
CQT	CNN	85.22	15.44
CQT	LCNN	85.90	15.22
CQT	ResNet	84.09	23.39

observed from Table 5.2 that CQT has high temporal resolution to be able to detect pop noise in affricate, fricative, and plosive sounds, whereas for STFT, the analysis window duration is fixed in the entire time-frequency plane and hence, it performs poorer than the proposed approach. On the other hand, production of nasal sound involves complete closure of the oral cavity and passing of quasi-periodic airflow puffs (from the vibrating vocal folds) to the nasal cavity due to lowering of the velum. Due to large volume of nasal cavity (than the oral cavity), nasal sound is dominated by low frequency resonance (nasal formant) and ESD. Furthermore, nasal formants have higher -3 dB bandwidth, since various energy losses are high as quasi-periodic airflow passes through complexly configured surface and thus, quickly damping its impulse response [42]. Thus, due to fixed duration of analysis window in STFT, it is difficult to detect pop noise event in the nasal sounds, whereas CQT is able to do it very effectively, due to its variable spectro-temporal resolution.

5.2.5.5 Inclusion of Replay Mechanism

The VLD can be viewed as another approach for SSD system, i.e., the pop noise detection in low frequency regions. Hence, in the framework of discussion related to Figure 5.2 in sub-Section 5.2.2.1, we extend this study by incorporating (simulated) replay mechanism in RP-A subset, which results in REP-A subset. This replay mechanism will enhance the characteristics of the spoof speech signal. We performed the experiments for classification of REP-A *vs.* RC-A to analyze the performance of the VLD system, when spoof speech signal consists of replay

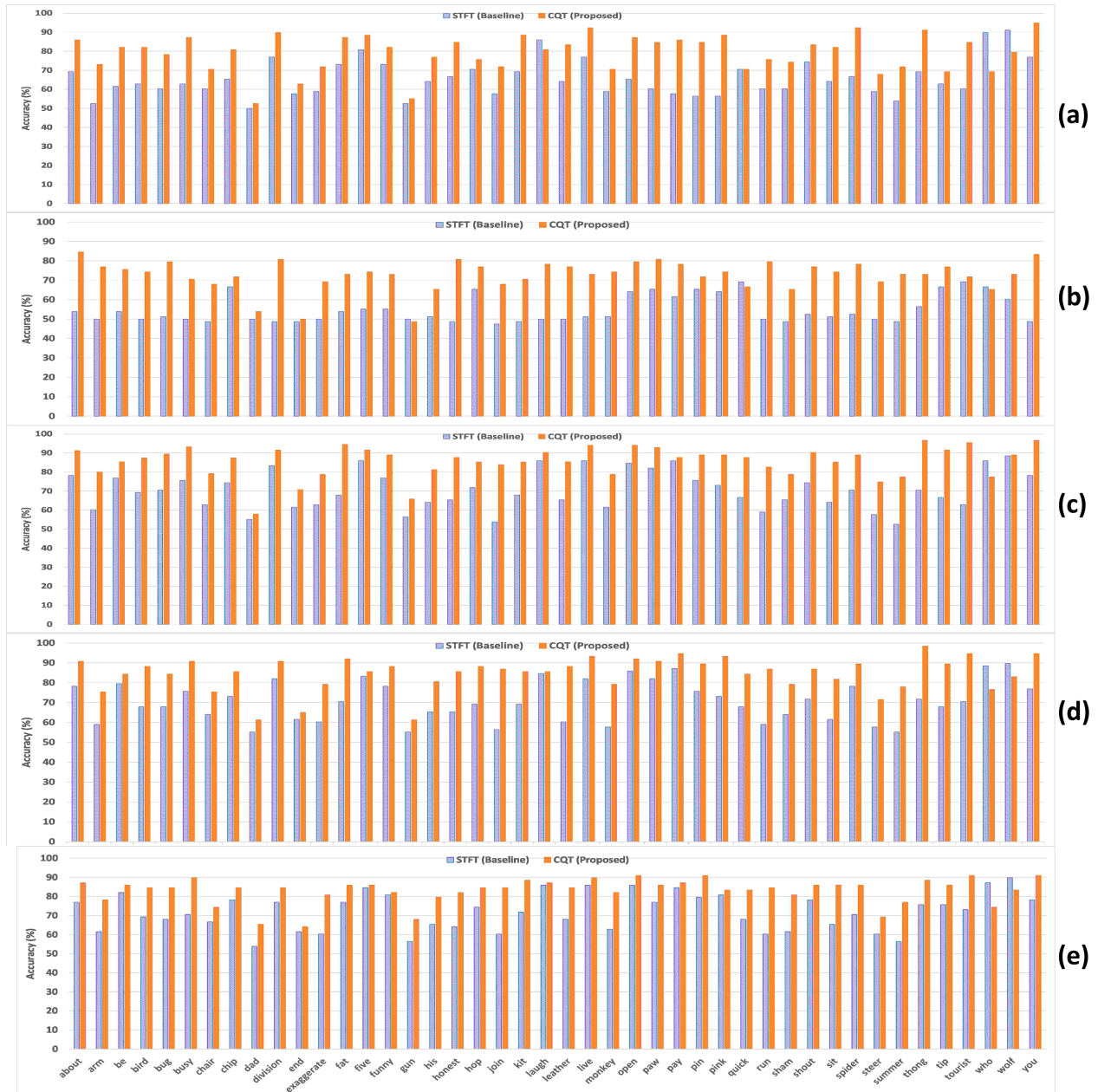


Figure 5.9: Comparison of Wordwise % Classification Accuracy on Dev set with (a) SVM, (b) GMM, (c) CNN, (d) LCNN, and (e) ResNet as Classifier for STFT (Baseline) and CQT (Proposed) Feature Set. After [15].



Figure 5.10: Comparison of Wordwise % Classification Accuracy on Eval set with (a) SVM, (b) GMM, (c) CNN (d) LCNN, and (e) ResNet as Classifier for STFT (Baseline) and CQT (Proposed) Feature Set. After [15].

Table 5.14: Comparison of Baseline *vs.* Proposed Approach for Different Types of Phonemes using Various Classifiers. After [15].

Feature set	Classifier	Average Accuracy Eval (%)			
		Affricate	Fricative	Plosive	Nasal
STFT	SVM	74.43	67.83	61	65.11
STFT	GMM	52.83	53.69	53.55	55.11
STFT	CNN	76	70.39	66.16	74.22
STFT	LCNN	76.5	70.24	65.77	72.22
STFT	ResNet	76.66	69.81	66.11	72
CQT	SVM	82.56	79.12	72.88	81.22
CQT	GMM	77.12	72.36	73.44	73.11
CQT	CNN	84.78	80.57	77.89	82.89
CQT	LCNN	86	83.06	81.88	85.44
CQT	ResNet	84.5	79.93	78.44	83.77

characteristics along with pop noise. In addition, experiments are performed for RP-A *vs.* REP-A classification to analyze the effect on performance of SSD task with *only* replay characteristics. Furthermore, we analyze the effect of the replay mechanism on the performance of SSD task *w.r.t.* variation in the frequency range of analysis and the distance between the genuine speaker and attacker’s microphone.

- **Effect of Distance Between the Genuine Speaker and Attacker’s Microphone**

In this sub-Section, we analyze the effect of distance between the genuine speaker and attacker’s microphone for RC-A *vs.* REP-A subset by inclusion of replay mechanism. From Table 5.15, it can be observed that for 30 cm distance, i.e., the scenario when the attacker’s microphone is close to the genuine speaker, the % classification accuracy obtained is lowest. This was as expected since when the microphone is close to the speaker, the effect of room acoustics during its replay will be minimum and hence, the replayed signal will be similar to the genuine signal (i.e., the distortions in the replayed signal will be less). As we increase the distance between speaker and microphone, the % classification accuracy also increases and is almost stable after distance of 50 *cm*. This may be due to the fact that as the distance between the speaker and microphone increases, the effect of room acoustics also increases while replaying it and hence, the replayed signal contained more distortions. We have considered the minimum distance as 30 cm due to the fact that in genuine speech recording (RC-A), the distance between speaker and microphone is kept 10 cm, and it is assumed that the attacker will be

Table 5.15: Effect of Varying Distance between Subject Speaker and Microphone on Performance (in % Classification Accuracy) of RC-A *vs.* REP-A Subset with SVM as a Classifier. After [15].

Feature Set	Distance	Accuracy (%)		EER (%)	
		Dev	Eval	Dev	Eval
STFT	30 cm	67.42	67.05	30.56	31.54
STFT	50 cm	65.24	65.68	34.37	34.61
STFT	70 cm	62.00	60.36	38.45	35.85
STFT	90 cm	64.48	63.77	35.56	36.11
CQT	30 cm	88.71	87.02	11.73	12.84
CQT	50 cm	93.01	91.21	6.8	8.67
CQT	70 cm	92.74	91.73	7.15	8.12
CQT	90 cm	92.45	91.71	7.44	8.14

at least at a distance of more than 10 cm.

- **Performance Evaluation *w.r.t.* Only Replay Mechanism**

In this sub-Section, we analyzed the performance of the effect of the *only* replay mechanism by performing the classification of RP-A *vs.* REP-A subset with SVM as a classifier. This analysis shows the significance of the pop noise in SSD task. Here, original RP-A subset is considered as genuine speech samples. Whereas, REP-A subset, which is created by embedding the replay mechanism into the original RP-A subset, is considered as spoof speech samples. This creates the similar scenario as that of the SSD task in ASVSpooof 2019 PA dataset, i.e., classification of the genuine *vs.* spoof speech signal without presence of pop noise. It can be clearly observed from Table 5.15 and Table 5.16 that the results obtained with pop noise inclusion shows much better performance than without pop noise scenario in genuine *vs.* SSD task.

- **Effect of Frequency Range (f_{max})**

In the proposed CQT-based algorithm, the optimal value of the f_{max} is selected as 30 Hz. However, the effect of additionally incorporated replay mechanism can exist for the entire frequency range. Hence, we extend the analysis by increasing the value of the f_{max} in the proposed CQT-based algorithm. For this set of experiments, SVM is utilized as a classifier. The distance between genuine speaker and attacker’s microphone is fixed at 50 cm as performance saturates after that. From Table 5.17, it can be observed that for the higher frequency range, the % classification accuracy increases and the corresponding % EER decreases. For frequency

Table 5.16: Effect of Varying Distance between Subject Speaker and Microphone on Performance (in % Classification Accuracy) of RP-A *vs.* REP-A Subset with SVM as a Classifier. After [15].

Feature Set	Distance	Accuracy (%)		EER (%)	
		Dev	Eval	Dev	Eval
STFT	30 cm	67.25	57.97	34.92	30.00
STFT	50 cm	56.09	52.47	42.84	41.24
STFT	70 cm	52.97	54.02	47.08	46.24
STFT	90 cm	54.17	53.76	44.56	44.41
CQT	30 cm	74.36	73.48	23.71	24.65
CQT	50 cm	80.80	78.74	18.23	20.90
CQT	70 cm	81.99	80.91	18.1	18.53
CQT	90 cm	82.66	82.56	17.22	17.13

range of 1000 Hz, accuracy is 98.72 % and 97.82 % for Dev and Eval set, respectively, and EER is as low as 0.48 % and 0.92 % for Dev and Eval set, respectively. This may be due to the fact that as we consider the higher frequency range, the effect of replay mechanism, i.e., the reverberation effect also escalates and hence, the classifier obtains more distinguished acoustic cues and does the classification task more effectively.

5.2.5.6 Performance Evaluation using ASVSpooof Challenge Datasets

As ASVSpooof challenge datasets are well known datasets in the voice anti-spoofing literature, we have extended the experiments using proposed CQT-based feature set on ASVSpooof 2019 PA and ASVSpooof 2017 version-2 datasets. The performance is compared with the CQCC feature set, which was utilized as a baseline feature set in ASVSpooof challenge campaigns. The results are reported in Table 5.18, and it can be observed that comparable performance is obtained on ASVSpooof 2019 PA dataset for the proposed CQT-based feature set as compared to the CQCC feature set. Whereas, proposed CQT-based feature set shows the poor performance on ASVSpooof 2017 version-2.0 dataset. The proposed feature set is designed to capture the low frequency spectral characteristics of the signal, in particular, 0-30 Hz frequency band. The spooof utterances in ASVSpooof 2019 PA dataset are simulated using a range of real replay devices and carefully controlled acoustic conditions. This *simulated* replay mechanism has well-behaved bandpass characteristics that distort the low and high frequency characteristics in the replay spooof signal. Hence, the proposed CQT-based feature set might succeed to capture the low frequency distortions introduced due to the replay mechanism and

Table 5.17: Results (in % Classification Accuracy and % EER) for RC-A *vs.* REP-A with Various Frequency Ranges. After [15].

Frequency Range	% Acc.		% EER	
	Dev	Eval	Dev	Eval
1-10 Hz	90.03	88.88	9.69	10.60
1-20 Hz	91.78	90.23	7.98	9.63
1-30 Hz	93.01	91.23	6.88	8.67
1-40 Hz	89.48	87.67	9.45	11.89
1-50 Hz	89.25	87.15	8.78	10.92
1-60 Hz	91.20	87.71	8.03	11.59
1-70 Hz	89.19	86.33	7.00	10.61
1-80 Hz	95.16	92.47	4.72	7.38
1-90 Hz	85.81	84.21	4.80	5.82
1-100 Hz	96.71	95.47	3.12	4.25
1-200 Hz	97.23	96.79	1.79	3.04
1-400 Hz	97.32	97.02	1.42	2.18
1-600 Hz	99.01	98.08	0.92	1.56
1-800 Hz	99.21	98.59	0.73	1.35
1-1000 Hz	98.72	97.82	0.48	0.92

Table 5.18: Results (in % EER) on ASVSpooF 2019 PA and ASVSpooF 2017 Version-2.0 Dataset using Proposed CQT-based Feature Set *vs.* CQCC (Challenge Baseline).

Feature Set	Dataset Used	Dev	Eval
Proposed CQT-based CQCC (The parameters and frequency range of CQT is as mentioned in this study)	ASVSpooF 2019	11.12	12.75
	PA scenario	9.87	11.04
	ASVSpooF 2017	41.05	43.27
	version-2	12.27	18.81

consequently producing the good performance using feature set that exploit characteristics of low frequency regions. On the other hand, ASVSpooF 2017 version-2.0 dataset consists of spooF speech utterances with *real* replay mechanism, whose frequency response would not be well-behaved bandpass in nature. This might be the reason that the proposed CQT-based feature set could show the good performance on ASVSpooF 2019 PA dataset and poor performance on ASVSpooF 2017 version-2 dataset.

5.3 Spectral Root Homomorphic Filtering-Based Features for Replay SSD in ASV and VAs

In this Section, we exploited homomorphic filtering-based approach for feature extraction. For homomorphic filtering-based approaches, the speech signal can be expressed as convolution of the glottal airflow (i.e., speech excitation source) with the impulse response of the vocal tract system [265]. A replayed speech signal can be considered as the convolution of natural speech with the impulse responses of recording, and playback devices as well as the acoustic environments [72]. Hence, for replay detection, the challenge is to estimate the characteristics of the extra convolved elements with genuine speech signal. A blind deconvolution approach could be used that requires a-priori knowledge of signal components. If one of the convolved components of the signal is known, then the other could be easily estimated using linear inverse filtering. LP analysis can be used to estimate the system function, with assumption that the all-pole LTI system convolves with a train of pulses or random noise [266,267]. If the general characteristics of one of the signal component is known, then homomorphic deconvolution system model can be used. In this case, either of logarithmic homomorphic deconvolution system (LHDS) or using spectral root homomorphic deconvolution system (SRHDS) could be used to perform deconvolution operation [268–270]. In LHDS, convolutionally-combined signals are mapped to additively combined signals, on which time-gating is applied for signal separation [271]. The time-gating in cepstral-domain is known as *liftering*. In SRHDS, convolutional vector space is mapped to another convolutional vector space, where signal components are more easily separable by liftering operation [66]. For replay detection on ASV and VAs, we employ SRHDS along with Mel filterbank to derive SRCC feature set [20]. Furthermore, the feature extraction algorithm can be independently applied to the magnitude and phase of the spectrum to give Magnitude-SRCC (MSRCC) and Phase-SRCC (PSRCC), respectively. The key novelty in proposed SRCC feature set is the use of power-law nonlinearity that does not depend critically on the input amplitude and thus, suitable for VAs that predominantly use far-field speech signal. Furthermore, we investigated the two approaches to systematically choose the optimum value of γ -parameter, which are explained in detail along with supporting results.

5.3.1 Speech Signal Modeling

Speech signal, $x(n)$ can be expressed as the convolution of glottal airflow (i.e., excitation source signal), $g(n)$, with the impulse response of vocal tract system, $v(n)$ [42], i.e.,

$$x(n) = g(n) * v(n), \quad (5.25)$$

where symbol '*' refers to the convolution operation. For our application, we refer $x(n)$ in eq. (5.25) as genuine speech signal. The glottal airflow is quasi-periodic (or impulse-like) in nature for voiced speech. It can be approximated as pulse-train for speech signal modeling. The replayed version of the genuine speech signal includes additional components, which are impulse responses of playback device $pd(n)$, playback environment $pe(n)$, recording device $rd(n)$, and recording environment $re(n)$ [72]. These components in the replay speech signal, $y(n)$ are convolutionally-combined with the genuine speech signal, $x(n)$, i.e.,

$$y(n) = x(n) * pd(n) * pe(n) * rd(n) * re(n). \quad (5.26)$$

Eq. (5.26) can be written as,

$$y(n) = x(n) * N(n) = g(n) * v(n) * N(n), \quad (5.27)$$

where $N(n) = pd(n) * pe(n) * rd(n) * re(n)$, and it is the overall impulse response that represents distortion in the genuine speech signal due to replay attack. One of the component in $x(n)$ and hence, in $y(n)$, is the glottal airflow $g(n)$, which is quasi-periodic in nature. As the characteristics of one signal, $g(n)$, is known, homomorphic signal processing techniques can be used to estimate the cepstrum for rest of the signal. To that effect, $v(n)$ in genuine and $v(n) * N(n)$ in spoof speech signal can be estimated. These estimated components can serve as discriminative acoustic cues for the SSD task.

5.3.2 Cepstrum Analysis: Logarithmic vs. Spectral Root

In evaluation of the logarithmic cepstrum, convolutionally-combined vector space, $x(n) = g(n) * v(n)$, is mapped to the additively combined vector space, $\hat{x}(n) = \hat{g}(n) + \hat{v}(n)$, such that the contribution of glottal airflow $g(n)$, and impulse response of vocal tract system, $v(n)$ can be distinctly observed [42, 270]. The logarithmic cepstrum is estimated as the inverse Fourier transform of the logarithm of

the Z-transform of the given signal $x(n)$, i.e.,

$$\hat{x}(n) = \mathcal{Z}^{-1}(\log(\mathcal{Z}(x(n)))) \tag{5.28}$$

where $\mathcal{Z}(\cdot)$ represents the Z-transform operator. Because of the transformation given in eq. (5.28), convolutional vector space is transformed to additive vector space [66]. This transformation takes place in such a way that the duration of the pulse-train, $\hat{g}(n)$ remains the same as that of $g(n)$, however, $\hat{v}(n)$ should get compressed (in quefrency-domain) than the $v(n)$ [270]. Here, $\hat{x}(n)$, $\hat{g}(n)$, and $\hat{v}(n)$ are referred to as logarithmic cepstrum of their corresponding time-domain signals, $x(n)$, $g(n)$, and $v(n)$, respectively. With similar analogy, cepstrum of the replay speech signal is given by [268,270],

$$\hat{y}(n) = \hat{x}(n) + \hat{N}(n) = \hat{g}(n) + \hat{v}(n) + \hat{N}(n). \tag{5.29}$$

Algorithm 6 Computing the Spectral Root Cepstrum. After [42].

1. $x(n) = g(n) * v(n)$,
 2. On applying Z-transform, $X(Z) = G(Z) \cdot V(Z)$,
 3. $X(Z)$ raised to γ power, $[X(Z)]^\gamma = [G(Z)]^\gamma \cdot [V(Z)]^\gamma$,
i.e., $\check{X}(Z) = \check{G}(Z) \cdot \check{V}(Z)$,
 4. Applying inverse Z-transform, $\check{x}(n) = \check{g}(n) * \check{v}(n)$,
 5. Perform the same operations on $y(n)$ to get $\check{y}(n)$.
-

Spectral root cepstrum is obtained by transforming the convolutional vector space, $x(n) = g(n) * v(n)$, to another convolutionally-combined vector space, $\check{x}(n) = \check{g}(n) * \check{v}(n)$, such that the elements are more easily separable in the transformed vector space than the earlier one. Here, $\check{x}(n)$, $\check{g}(n)$, and $\check{v}(n)$ represents spectral root cepstrum of the signals $x(n)$, $g(n)$, and $v(n)$, respectively. Here, logarithmic non-linearity in LHDS is replaced by *power-law nonlinearity*, $(\cdot)^\gamma$.

Spectral root cepstrum for signals, $x(n)$ and $y(n)$ are expressed as follows:

$$\check{x}(n) = \check{g}(n) * \check{v}(n), \tag{5.30}$$

$$\check{y}(n) = \check{x}(n) * \check{N}(n) = \check{g}(n) * \check{v}(n) * \check{N}(n). \tag{5.31}$$

The steps for computing the spectral root cepstrum are given in Algorithm 6. In general, pole-zero sequence of $v(n)$ is an infinitely long sequence [66]. Hence, $\check{v}(n)$ will also be infinitely long. In this situation, we select ' γ ' to maximally compress $\check{v}(n)$ such that it has the smallest energy concentration in the low-time region, where $-1 \leq \gamma \leq 1$. The appropriate value of the ' γ ' depends upon the pole-zero

combination of the system function. To understand the appropriate choice of γ depending upon pole-zero location of the system function, let us assume $g(n)$ is a periodic train of pulses then, according to the computation given in Algorithm 6, $\check{g}(n)$ will also be the periodic train of pulses with similar duration [42, 66]. In particular,

$$g(n) = \delta(n) + \beta\delta(n - N), \quad 0 < \beta < 1, \quad (5.32)$$

where $\delta(n)$ represents the unit impulse function, and N represents the spacing between two impulses. If $v(n)$ is the all-pole sequence of order q and $q < N$, then:

$$G(z) = 1 + \beta z^{-N}. \quad (5.33)$$

Suppose we form the spectral root cepstrum of $x(n)$ with $\gamma = -1$. Then, using the Taylor series expansion for $\frac{1}{1+\beta z^{-1}}$ and replacing z by z^N , it can be seen that the inverse z -transform of $G^{-1}(z)$ is an impulse train with impulses spaced by N samples. Also, $V^{-1}(z)$ is all-zero, since $V(z)$ is all-pole, so that $\check{v}(n)$ is a q -point sequence. Because $q < N$, $v(n)$ can be deconvolved from $x(n)$ by inverting $X(z)$ to obtain $X^{-1}(z)$, and liftering $g^{-1}(n)$, the inverse z -transform of $G^{-1}(z)$, using a right-sided lifter of q samples. Similarly, for all-zero sequence, $\gamma = 1$ will be the desirable value to efficiently extract $\check{v}(n)$. However, in practical case, the value of γ will vary between -1 to +1. If the number of poles are larger than the zeros, then the desired value of γ should be negative and vice-versa.

The energy concentration of the first n points of $\check{v}(n)$ relative to its total energy is given by [66]:

$$d(n) = \frac{\sum_{k=1}^n |\check{v}(k)|^2}{\sum_{k=1}^C |\check{v}(k)|^2}, \quad (5.34)$$

where n is a number of samples in time-gating [66], and C corresponds to the total number of cepstral coefficients. For our application, we select γ , and number of samples (n in eq. (5.34)) for time-gating such that it can discriminate between $\check{v}(n)$ (eq. (5.30)) in genuine speech samples, and $\check{v}(n) * \check{N}(n)$ (eq. (5.31)) in spoof speech samples. The application of eq. (5.34) to select the appropriate value of γ , is discussed in Section 5.3.5.

5.3.3 Proposed SRCC Feature Set

In this study, we proposed to use the SRCC feature set, which uses the triangular Mel filterbank along with power-law nonlinearity as shown in Figure 5.11. Furthermore, the processing is performed on the magnitude and phase part of the spectrum to produce MSRCC and PSRCC, respectively. Windowing is per-

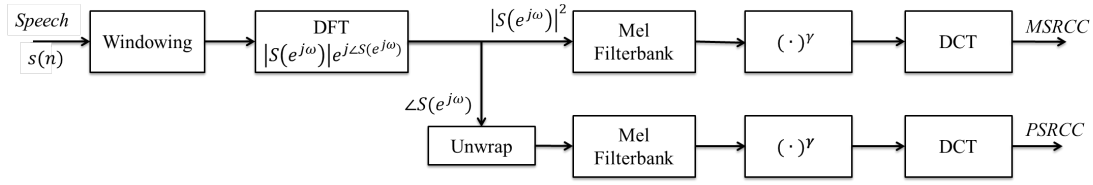


Figure 5.11: Functional Block Diagram of SRCC (MSRCC and PSRCC) Feature Set Extraction. After [20].

formed on the input speech signal with optimum window and hopping length as reported in speech signal processing literature. The q^{th} cepstral coefficient is extracted using magnitude spectrum to give MSRCC as [19]:

$$MSRCC(q) = \sum_{m=1}^M (MFM(m))^\gamma \cos \left[\frac{q(m - \frac{1}{2})\pi}{M} \right], \quad (5.35)$$

where the Mel Frequency Magnitude (MFM) spectrum is defined as:

$$MFM(m) = \sum_{k=1}^K |X(k)|H_m(k), \quad (5.36)$$

where $X(k)$ represents k -point DFT of the signal, $x(n)$, $H_m(k)$ is the m^{th} Mel-scaled bandpass filter. Root Spectrum in Mel Scale (RSMS) is given by:

$$RSMS(m) = (MFM(m))^\gamma. \quad (5.37)$$

The value of γ is chosen so as to have maximum distinction between genuine and spoof utterances of the ReMASC dataset. Furthermore, PSRCC is derived using unwrapped phase as:

$$PSRCC(q) = \sum_{m=1}^M (MFP(m))^\gamma \cos \left[\frac{q(m - \frac{1}{2})\pi}{M} \right], \quad (5.38)$$

where the Mel Frequency Phase (MFP) spectrum is defined as:

$$MFP(m) = \sum_{k=1}^K \angle X(k)H_m(k), \quad (5.39)$$

where $\angle X(k)$ represents the unwrapped phase of the $X(k)$.

5.3.4 Experimental Setup

In this study, we utilized the ReMASC dataset to build the CMs against the replay attack for VAs and ASV system [9]. The dataset configuration used for the experiment for ReMASC and ASVspoof 2017 challenge datasets is shown in Table 3.15 and Table 3.3, respectively, of chapter 3. The performance of the MFCC, LFCC, and CQCC feature sets is explored along with the proposed MSRCC, PSRCC, and RSMS feature representations. GMM, CNN, and LCNN are utilized as classifiers along with % EER as evaluation metric for various experiments in this work. Score-level fusion is performed to obtain the possible complementary information in various SSD systems.

5.3.5 Experimental Results on ReMASC Dataset

For the given dataset, we have approximated the value of the γ with the help of eq. (5.34). Initially, we have chosen three different values of γ , i.e., -0.9, 0.1, and 0.9. The MSRCC features are extracted as explained in sub-Section 5.3.3. We computed the energy concentration over 4000 genuine and spoof samples for $n = 13$ in eq. (5.34) and averaged over all the utterances. For $\gamma = 0.9$ and -0.9 , we obtained 81% energy concentration for $n = 13$ coefficients. Whereas, approximately 87 % energy is preserved for the same value of n , when γ value is set to 0.1.

Furthermore, we observed the spectrogram obtained by varying the value of γ in eq. (5.37). Figure 5.12 shows the spectrogram of the genuine and spoof speech signal in Panel-I and Panel-II, respectively. Figure 5.12 (a), (b), (c), and (d) are obtained for the values of γ as 0.9, -0.9 , 0.1, and -0.1 , respectively. It can be observed that maximal information of the input speech signals is captured in the spectrogram with $\gamma = 0.1$, as seen in Figure 5.12(c), where the spectral contents of the speech signal seems to be well enhanced as compared to the Figure 5.12(a), Figure 5.12(b), and Figure 5.12(d). In [66], the performance of the SRHDS is also evaluated *w.r.t.* varying the number of poles and zeros in the system function. The value of γ varies between -1 to 1. If system function consists of only poles then $d(n)$ in eq. (5.34) is maximized for $\gamma = -1$. However, for all-zero system, $\gamma = 1$ is the appropriate choice [42, 66]. For our case, $d(n)$ is maximized for $\gamma = 1/11$ (obtained after further fine-tuning), which suggests that the system function has more zeros than the poles. With this analysis, we have chosen the value of γ to be 1/11 for our SSD system design.

Furthermore, spectrographic analysis for RSMS *vs.* CQT is shown in Figure 5.13. Panel I shows the RSMS for genuine and spoof speech signal in Figure 5.13(a)

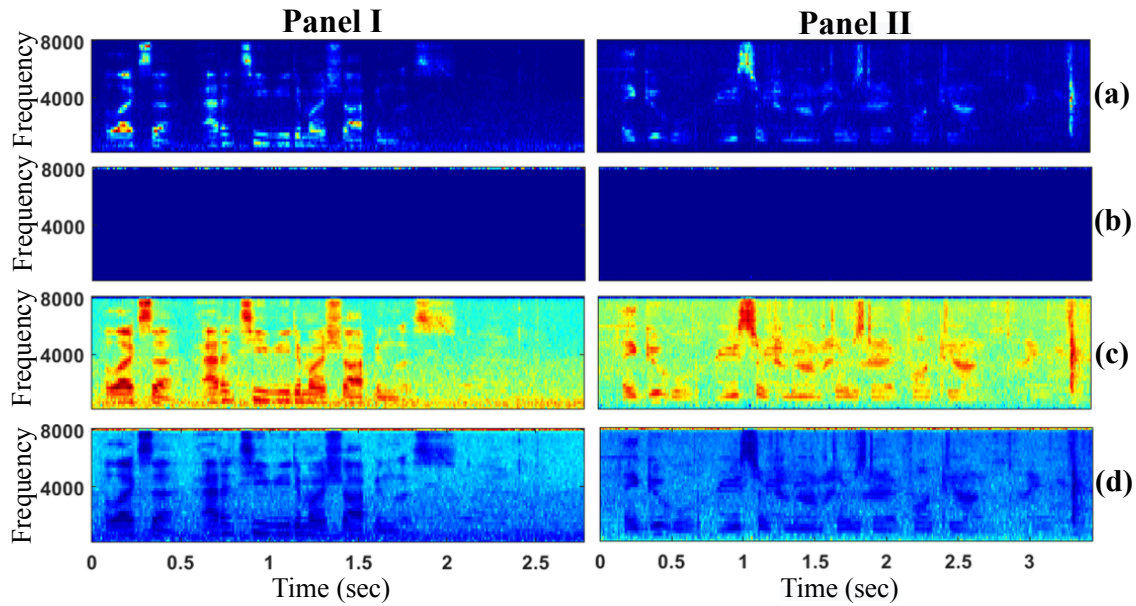


Figure 5.12: Panel-I and Panel-II Consists of Spectrogram of Genuine *vs.* Spoof Speech Signal, Respectively. Figure 5.12(a) Shows Spectrogram of the Speech Signal as Given in eq. (5.37) for $\gamma = 0.9$. Whereas, Figure 5.12(b), Figure 5.12(c), and Figure 5.12(d) Shows the Spectrogram for $\gamma = -0.9, 0.1$, and -0.1 , Respectively. After [19].

and Figure 5.13(c), respectively. Figure 5.13(b) and Figure 5.13(d) in Panel II shows CQT-gram for genuine and spoof speech signal, respectively. It can be observed that the highlighted region in RSMS is highly *discriminative* than the CQT-gram. It may be due to the fact that the choice of appropriate γ helps by preserving the maximum concentration of signal's energy and hence, its behaviour is more profoundly observed. It can be observed from the Figure 5.13 that a spoof signal has lesser spectral energy in the high frequency region as compared to its genuine counterpart, which may be due to energy decay because of replay configuration. Thus, being able to capture the behaviour of the signal in the high frequency region allows RSMS to distinguish genuine and spoof utterances effectively.

Experiments are performed by varying the value of γ for MSRCC feature set along with GMM classifier, and results in % EER are shown in Table 5.19. It is observed that the $\gamma > 0$ gives better results than $\gamma < 0$, which is exactly opposite to what has been reported for speech synthesis application (Chapter 6 and Chapter 9 of [42]). In particular, $\gamma = -1/3$ gave better results in listening test than $\gamma = 1/3$. This is because $\gamma < 0$ emphasize the pole structure (i.e., formants), whereas $\gamma > 0$ emphasize the zeros (i.e., valleys) in the spectrum. For replay speech, due to bandpass characteristics of the replay mechanism, the spectrum is expected to decay faster, which is essentially encoded in the valleys in the spec-

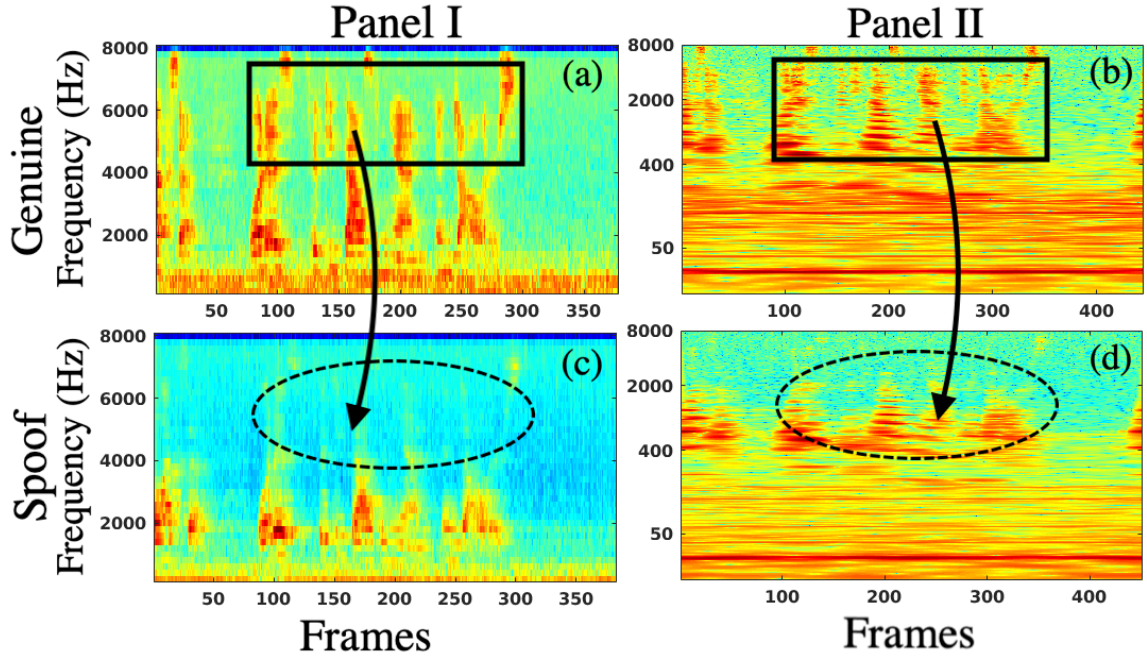


Figure 5.13: Plot of RMS (Panel I) *vs.* CQT-gram (Panel II) Feature Sets : (a), (b) for Genuine Speech Signal, and (c), (d) for Spoofed Speech Signal. After [19].

trum. Thus, $\gamma > 0$ is able to emphasize this information better than $\gamma < 0$ and hence, gives the relatively better results.

Table 5.19: Variation in % EER *w.r.t.* γ Value for MSRCC Feature Set. After [19].

γ	-1	-1/2	-1/3	-1/5	-1/7	-1/9	-1/11	-1/13	-1/15
Dev	38.78	30.64	27.92	23.04	22.69	21.59	20.94	20.46	20.25
Eval	33.33	26.17	23.74	21.59	21.38	21.12	20.74	20.40	20.85
γ	1/15	1/13	1/11	1/9	1/7	1/5	1/3	1/2	1
Dev	21.56	20.19	19.27	20.28	21.13	21.28	23.78	24.67	32.36
Eval	19.23	18.20	16.16	19.90	20.79	22.01	24.65	25.75	31.30

Results obtained are shown in Table 5.20 with % EER as a performance metric. For MSRCC-GMM (A) system, we obtained the absolute reduction in EER of 1.3% and 7.05% on the Dev and Eval sets, respectively, in comparison with the baseline CQCC-GMM system. In addition, it can be observed from Table 5.20 that, RMS-LCNN (B) system performs superior to the CQT-gram-LCNN system. This validates the efficacy of the proposed spectral root-based feature sets, irrespective of the classifier. The improved results on Eval set using MSRCC feature set show its generalization capability. The performance of the MSRCC is also compared with the Power Normalized Cepstral Coefficients (PNCC) and RASTA-PLP feature sets, which are also power-law nonlinearity-based features. These feature

Table 5.20: Results (in % EER) on ReMASC Dataset using Various Feature Sets. After [19].

SSD System	Dev	Eval
CQCC-GMM	20.57	23.31
LFCC-GMM	28.89	26.31
MFCC-GMM	36.43	31.53
PNCC-GMM	22.29	25.23
RASTA-PLP - GMM	26.25	29.20
MSRCC-GMM (A)	19.27	16.26
CQT-gram-LCNN	13.20	15.14
RSMS-LCNN (B)	13.75	12.24
A + B	11.39	11.84

‘+’ denotes score-level fusion.

sets are developed to design noise-robust speech recognition system [272]. Because of their noise-robustness property, they may fail to detect the distortions in replayed spoof speech signals. We found that PNCC also works better for the positive values of γ near zero and results in Table 5.20 are obtained with power-law nonlinearity exponent set to $\gamma = \frac{1}{11}$. Whereas RASTA-PLP shows quite consistent performance with variation of this exponent, however, we obtained somewhat better results, when it is set to $\gamma = \frac{1}{10}$. We also used LFCC and MFCC feature sets in our experiments. Both MFCC and LFCC feature sets perform poorer than the CQCC feature set. Our primary system shows the absolute reduction in EER of 6.82 % and 11.07 % on Dev and Eval sets, respectively, compared to the baseline system. The fusion of the two SSD systems A and B, with fusion parameter $\alpha = 0.33$, gives much better performance, suggesting both of these systems capture complementary information. Experiments performed on environment-independent case produces relatively better results than the CQCC-GMM system reported in [9] (i.e., 31.60 %, 29.78 %, 26.40 %, and 36.23 % EER in four successive environments).

5.3.6 Experimental Results on ASVSpooF 2017 Dataset

In this Section, we describe the development of SSD using the proposed feature sets. The experiments are performed on the ASVSpooF 2017 challenge database. The full database contains three subsets: training, Dev, and Eval set. The details of ASVSpooF 2017 is given in [165, 273]. Initially, experiments are performed to optimize the feature and model parameters, namely, the value of γ , frequency range

Table 5.21: Results (in % EER) on ASVSpooof 2017 Dataset. After [20].

SSD System	Dev	Eval
CQCC-GMM (Baseline) [255]	12.11	29.18
MFCC-GMM (Baseline) [255]	11.21	31.30
MSRCC-GMM	8.53	18.61
PSRCC-GMM	35.53	24.35
MSRCC + PSRCC (GMM)	6.58	10.65
MSRCC-CNN	3.05	24.84
PSRCC-CNN	36.21	26.81
MSRCC + PSRCC (CNN)	2.63	17.76

in the spectrum, dimension of feature vector, and number of mixture components in the GMM. The experiments are carried for various values of the γ ranging from -1 to +1 for MSRCC and PSRCC feature sets. It was observed that the optimum performance is observed for $\gamma = -1/7$. Furthermore, experiments are also performed by selecting various frequency range for both MSRCC and PSRCC feature sets. PSRCC feature set performs better for the entire frequency range, whereas MSRCC performs better for the 6-8 kHz spectral feature representations. Furthermore, number of feature dimension and number of mixtures in GMM are selected empirically for the optimum performance as 13 and 512, respectively. The experiments are also performed using CNN as a classifier. The experimental results are as shown in Table 5.21. The CQCC-GMM and MFCC-GMM was the baseline systems for ASVSpooof 2017 challenge. The proposed MSRCC and PSRCC feature sets performed better over the baseline results for both GMM and CNN classifiers. Furthermore, the score-level fusion of the MSRCC-GMM and PSRCC-GMM system shows the significant amount of improvement in the performance. It shows that the complementary information is present in magnitude and phase representation for the replay SSD task. The similar results are shown in Figure 5.14 using the DET plots.

5.4 Chapter Summary

In this study, we exploited the CQT-based algorithm to detect the liveness in the genuine speaker by using the pop noise as a discriminative acoustic cue. The experiments are performed on the recently released POCO dataset. The results of the proposed approach are compared against the baseline, where feature sets are derived from the traditional STFT. The spectrographic analysis for genuine

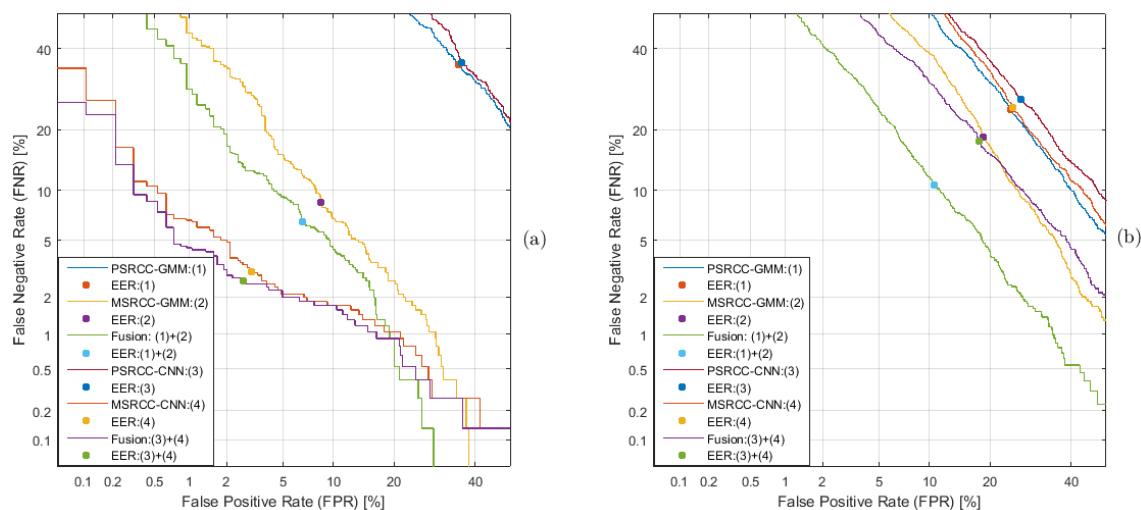


Figure 5.14: DET Curves for (a) Dev Set and (b) Eval Set of ASVSpooof 2017 Challenge Dataset. After [20].

(live) *vs.* spoof speech was performed, which showed that the pop noise is emphasized in a much better way in CQT-based spectrogram than it's STFT counterpart. The VLD systems for proposed CQT-based algorithm *vs.* STFT-based baseline are developed using various classifiers, namely, SVM, GMM, CNN, LCNN, and ResNet. The performance of the VLD system is evaluated using two performance evaluation metrics, namely, % classification accuracy and % EER. It was found that the proposed CQT-based algorithm performs better over STFT-based baseline algorithm, for all the four classifiers. Furthermore, experiments are performed using various analysis windows, namely, Hann, Hamming, and Gaussian. Among these windows, Gaussian window gave relatively better performance over hann and Hamming window functions and hence, we reported remaining results using Gaussian window. Relatively better performance using Gaussian window, might be because of the fact that Gaussian window achieves lower bound for Heisenberg's box corresponding to uncertainty principle in signal processing framework [200]. The relatively best performance is obtained by CQT-based algorithm along with LCNN architecture among all the VLD systems considered in this study. However, there is still a scope for further improvement by utilizing more efficient pop noise detection methods (improved signal processing or probabilistic approaches) and sophisticated deep learning architectures, in particular, by exploiting various loss functions in CNN, LCNN, and ResNet [15].

In addition, we have analyzed the effect of the addition of replay mechanism by generating simulated replay. In order to generate simulated replay, the geometrical acoustics is generated by using image-source model. We have analyzed the effect of pop noise with variation in frequency range. It was found that the

effect of the pop noise is significantly observed below 30 Hz. Furthermore, it is observed that after inclusion of simulated replay mechanism in pop noise detection framework, the classification accuracy is increased with increase in frequency range. This is due to the addition of replay mechanism, the reverberation effect is introduced in replayed speech and hence, a significant difference is obtained at the acoustic-level between genuine (live) *vs.* replayed speech. However, pop noise can be captured only from a short distance and hence, this approach of replay SSD is effective only when the distance between microphone and speaker is less for genuine speech recordings. Moreover, it is assumed in the dataset that the distance between attacker's recording device and genuine speaker is large and hence, the pop noise does not get captured by the recording device. The addition of the pop noise in the spoof speech utterance can be easily performed. However, the proposed approach did not address the issue of artificially added pop noise in spoof speech signal, which can be easily added at the arbitrary locations in the utterance. Thus, the VLD system can be further modified to detect the pop noise at pop noise-specific phonemes and improve the security for the ASV system, which remains an open research problem.

Furthermore, the SRCC feature set is employed for replay SSD task. We investigated physics of replay attack and spectral root cepstrum, where logarithmic nonlinearity in MFCCs is replaced by power-law nonlinearity for replay SSD in the context of VAs. For power-law nonlinearity, dynamic behavior of the output does not depend critically on the input amplitude. A proper choice of γ in SRCC feature extraction plays a vital role in deconvolving the input signal. The selected γ value also pointed out that this system possess more zeros than the poles. The experiments are performed on ASVSpooF 2017 and ReMASC datasets using MSRCC and PSRCC feature sets. For ASVSpooF 2017 dataset, MSRCC and PSRCC feature sets extract complementary information and hence, their score-level fusion produces significant improvement in results. However, ReMASC dataset shows the better performance with MSRCC feature set alone.

The next chapter discusses the contribution of this thesis work in the other work on anti-spoofing and other speech technology applications, namely, classification of severity-level of dysarthric speech, and classification of normal *vs.* pathological infant cries.

CHAPTER 6

Other Applications

6.1 Introduction

In¹ the earlier chapters, the proposed handcrafted feature sets for anti-spoofing task were discussed. During this work, various feature sets and the classifier architectures for anti-spoofing task were also applied to other applications, such as severity-level classification of dysarthric speech and classification of normal *vs.* pathological infant cries. This chapter presents the major work performed on allied paths of feature development on anti-spoofing or the other applications of speech technologies. To that effect, this chapter deals with following *four* major components:

1. Significance of feature normalization methods, such as CMVN as a double-

¹This Chapter is based on the following publications:

- **Ankur T. Patil**, and Hemant A. Patil, "Significance of CMVN for Replay Spoof Detection," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Auckland, New Zealand, Dec. 7-10, 2020, pp. 532-537.
- Piyushkumar Chodingala, Shreya Chaturvedi, **Ankur T. Patil**, and Hemant A. Patil, "Robustness of DAS Beamformer Over MVDR for Replay Attack Detection On Voice Assistants," accepted in IEEE International Conference on Signal Processing and Communications (SPCOM)-2022, Bangalore, India, July 11-15, 2022.
- Siddhant Gupta, **Ankur T. Patil**, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A. Patil, and Rodrigo C. Guido, "Residual Neural Network Precisely Quantifies Dysarthria Severity-level based on Short-duration Speech Segments", in Neural Networks, Elsevier, 139(2021): 105-117.
- Hemant A. Patil, **Ankur T. Patil**, Aastha Kachhi, "Constant Q Cepstral Coefficients for Classification of Normal *vs.* Pathological Cry," accepted in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, May 7-13, 2022, pp 7392–7396.
- **Ankur T. Patil**, Aastha Kachhi, Hemant A. Patil, "Subband Teager Energy Representations for Infant Cry Analysis and Classification," accepted in European Signal Processing Conference (EUSIPCO)-2022 Belgrade, Serbia, August 29 - Sept. 2, 2022.

- edged sword for anti-spoofing;
- 2. Analysis of Delay and Sum (DAS) *vs.* MVDR beamforming techniques for anti-spoofing in VAs;
- 3. Severity-level classification of dysarthric speech;
- 4. Infant cry classification.

The subsequent sections of this chapter consists of brief explanation of the above components of the work.

6.2 Significance of CMVN for Replay Spoof Detection

6.2.1 Motivation

Feature normalization techniques have been used in various speech applications, such as automatic speech and speaker recognition, to improve the performance of the systems [274–279]. The literature includes several forms of normalization techniques, which includes normalization w.r.t. n^{th} order expectation of random variable X for each dimension. The first and second-order expectations are known as mean and variance, respectively. If normalization is applied on the cepstral feature representation based on mean and variance, then it is known as Cepstral Mean and Variance Normalization (CMVN). However, if we consider only mean value for normalization, then it is called as CMN or Cepstral Mean Subtraction (CMS) [280, 281]. The use of normalization techniques in ASR for environmental mismatch conditions is well known in the literature [275, 276, 279]. Further, these normalization methods are also used for speaker recognition [281–283].

The replay spoof speech signal is formed by convolving the genuine version of the speech sample with the impulse responses of the recording and replay environments and devices. In SSD task, we need to identify this additional transmission channel effects present in spoof speech signal. The application of the CMVN/CMN to the speech and speaker recognition system suppresses the transmission channel effects. Hence, its use in SSD task seems to be counter-intuitive. However, among the many CM systems developed on ASVspoof-2017 dataset, it is observed that CMVN/CMN has been effectively utilized for the replay SSD task to give significant improvement in the performance [3, 99, 103, 105–107, 195, 284,

285]. This contradictory results motivated us for further investigation over applicability of the CMVN. We performed experiments for environment-independent and environment-dependent scenarios. Furthermore, probability density function (*pdfs*) are estimated over several dimensions of feature representations.

6.2.2 Cepstral Mean and Variance Normalization (CMVN)

CMN was initially proposed to eliminate the transmission channel distortions that are introduced into the signal by convolving the signal with the impulse response of the transmission channel. In cepstral-domain, convolutional vector space is mapped to the additive vector space [42,268]. The CMN estimates the mean along every dimension of the cepstral feature representation of the speech samples and this mean value is subtracted from the corresponding dimension to transform the feature representation to zero-mean. Whereas, the CMVN transforms each cepstral feature representation of the speech sample to zero-mean and unit-variance. Mean and variance can be estimated for a segment of the utterance to reduce the latency period [279].

Let x_t denote the d -dimensional feature vector at the frame index t of the utterance, and $x_t(i)$ represent the i^{th} component of x_t . The speech utterance is passed through frame-blocking, denoted as $X = [x_1, x_2, \dots, x_T]$, where T denote the number of speech frames. The mean and variance are estimated for every dimension in maximum likelihood (ML) framework as [286]:

$$\mu_{ML}(i) = \frac{1}{T} \sum_{t=1}^T x_t(i), \quad 1 \leq i \leq d, \quad (6.1)$$

$$\sigma_{ML}^2(i) = \frac{1}{T-1} \sum_{t=1}^T [x_t(i) - \mu_{ML}(i)]^2, \quad 1 \leq i \leq d, \quad (6.2)$$

where μ_{ML} and σ_{ML} corresponds to mean and variance values, estimated in ML framework. The CMVN is applied to obtain normalized cepstrum of the frame as:

$$\hat{x}_t(i) = \frac{x_t(i) - \mu_{ML}(i)}{\sigma_{ML}(i)}, \quad 1 \leq t \leq T, \quad 1 \leq i \leq d. \quad (6.3)$$

To visualize the effect of the CMN and CMVN, we generated the data samples with the help of two random variables from the normal distributions, $\mathcal{N}(4,4)$, and $\mathcal{N}(2,0.25)$. The scatter plot of the generated data samples is shown in Figure 6.1(a). Further, Figure 6.1(b) and Figure 6.1(c) shows the scatter plot for CMN and CMVN normalized data samples, respectively. For CMN data samples, it can be

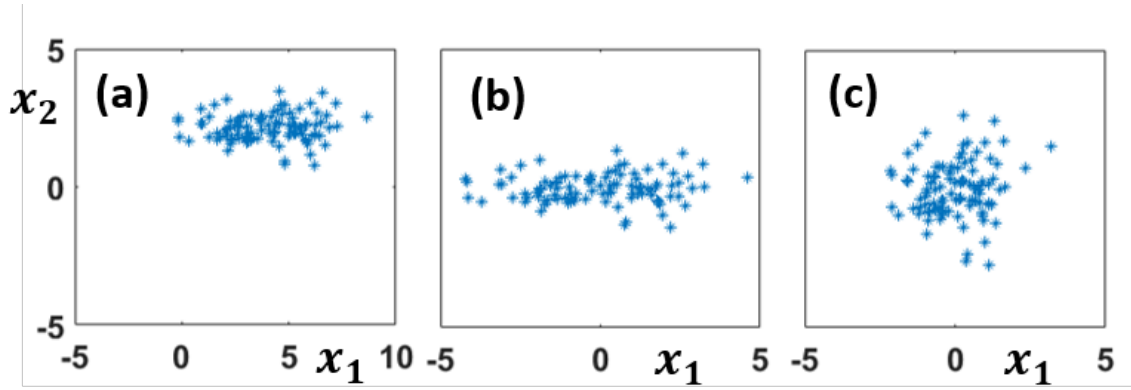


Figure 6.1: Scatter Plot for (a) the Unnormalized Data, (b) with CMN, and (c) CMVN. $X = [x_1 \ x_2]$ Denotes the Samples Drawn from the Bivariate Gaussian Distribution. Ytick Values of Figure 6.1(b) and Figure 6.1(c) are the Same as that of Figure 6.1(a). After [21].

observed that the data samples are centered around the origin, and variance is maintained the same as that of the original data samples, however, with CMVN, the mean and variance are normalized. The spread along both the axes is maintained at unity variance in CMVN as shown in Figure 6.1(c).

Similar kind of observations regarding feature normalization can be seen in Figure 6.2, which shows the scatter plots for the first two dimensions (D) of the CQCC feature set for genuine *vs.* two spoof speech signals. The data samples for the genuine speech are shown by red '*' symbol, whereas spoof speech data samples for balcony and studio environments are shown by green and blue '*' symbol, respectively. The CQCC feature extraction and dataset details are discussed in Chapter 3. Figure 6.2(a), Figure 6.2(b), and Figure 6.2(c) shows the scatter plots for original features (initial 2- D), it's CMN, and CMVN normalized versions, respectively. Again, feature representation with CMN is centered around origin with original feature representation variance, whereas feature representation with CMVN is zero-centered with unity variance. The other intuition from this scatter plot is discussed in the next Sections.

6.2.3 Replay Speech Signal Modelling and CMVN

Using source-filter model, the speech signal, $s(n)$ is modelled as the convolution of the glottal airflow, $g(n)$ with the impulse response of the vocal tract system, $v(n)$, i.e.,

$$s(n) = g(n) * v(n). \quad (6.4)$$

In many speech signal processing applications, the speech signal is represented in the cepstral-domain. The cepstral representation of the speech signal is obtained as the inverse Fourier transform of the logarithm of the spectrum of the speech signal. This transformation maps the convolutionally-combined vectors to additively combined vectors [66, 265, 268–271]. Let $\hat{s}(n)$, $\hat{g}(n)$, and $\hat{v}(n)$ represents the cepstrum of the speech signal, glottal airflow, and vocal tract system, respectively. Cepstral representation of the speech signal in eq. (6.4) is given as:

$$\hat{s}(n) = \hat{g}(n) + \hat{v}(n). \quad (6.5)$$

In SSD framework, the signal $s(n)$ is treated as genuine speech signal. The effect of the distortion due to replay mechanism on genuine speech signal, can be modelled by linear filtering. In the replay mechanism, the genuine signal is recorded, and again replayed back. In this process, the genuine signal is distorted by the impulse responses of the recording environment, $a(n)$, recording device, $b(n)$, playback device, $c(n)$, and playback environment, $d(n)$, respectively. By linear filter theory, the replayed speech is referred to as convolution of the genuine speech signal with this additional components, i.e.,

$$r(n) = s(n) * a(n) * b(n) * c(n) * d(n). \quad (6.6)$$

Let all these additional elements ($a(n)$, $b(n)$, $c(n)$, and $d(n)$) contribute to the overall impulse response of the replay mechanism system, $h(n)$ (i.e., $h(n) = a(n) * b(n) * c(n) * d(n)$), which will distort the genuine speech signal. Hence, the replayed signal is modeled as:

$$r(n) = s(n) * h(n). \quad (6.7)$$

In the cepstral-domain, eq. (6.7) is written as:

$$\hat{r}(n) = \hat{s}(n) + \hat{h}(n). \quad (6.8)$$

If an utterance consists of T number of speech frames, then the cepstrum for each frame can be written as [287, 288]:

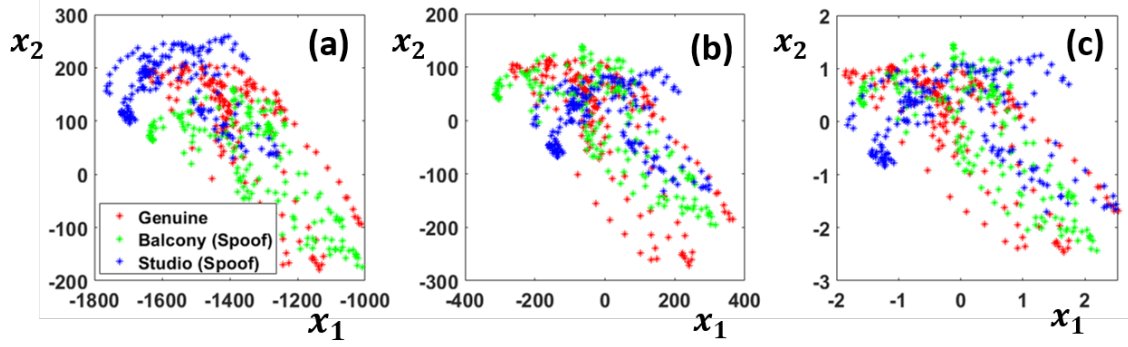


Figure 6.2: Scatter Plot for (a) the Unnormalized Data, (b) with CMN, and (c) CMVN. $X = [x_1 \ x_2]$ Denotes the First and Second Dimension of CQCC Feature Vector. Legends of Figure 6.2(b) and Figure 6.2(c) are the Same as That of Figure 6.2(a). After [21].

$$\begin{aligned}
 \hat{r}_1(n) &= \hat{s}_1(n) + \hat{h}(n), \\
 \hat{r}_2(n) &= \hat{s}_2(n) + \hat{h}(n), \\
 &\vdots \\
 \hat{r}_T(n) &= \hat{s}_T(n) + \hat{h}(n).
 \end{aligned} \tag{6.9}$$

Taking average over the T frames, we get,

$$\frac{1}{T} \sum_{t=1}^T \hat{r}_t(n) = \frac{1}{T} \sum_{t=1}^T \hat{s}_t(n) + \hat{h}(n). \tag{6.10}$$

Here, we modeled the effect of distortion by linear filter approach. Then, the distortion from the observed signal is removed by the inverse filtering. The cepstrum is computed as the inverse Fourier transform of the logarithm of the Fourier transform. Hence, the effect of the distortion can be removed (at least suppressed) by subtracting the characteristics of the distortion filter from the cepstrum of the observed signal. In eq. (6.10), the cepstrum of the distortion filter, $\hat{h}(n)$, can be subtracted to obtain the distortionless signal. Let us assume that the genuine speech signal is the zero-mean process. Then,

$$\hat{h}(n) = \frac{1}{T} \sum_{t=1}^T \hat{r}_t(n) = c_\mu. \tag{6.11}$$

Then, CMN is supposed to remove or at least suppresses the effect of the distortion from the replayed speech signal. Inevitably, the average over the cepstral coefficients include the speech and speaker-related information, and the effect of

the transmission channel distortion.

6.2.4 Experimental Setup

In this study, we aim to investigate the application of the CMVN for spoof detection capability over both environment-independent and environment-dependent cases for ASVSpooF 2017 and ASVSpooF 2019 datasets. In environment-independent case, the target environment is unseen by the defense model. To perform the experiments on environment-independent case, the same statistical distribution of the speech samples as provided by the ASVSpooF-2017 and -2019 (PA scenario) challenge organizers are used, which is shown in Table 3.3 and Table 3.5, respectively. For environment-dependent case, the target environment is seen by the defense model. In this case, training and testing is performed on each individual environment in the corresponding dataset. The distribution of the number of spoof speech utterances for each environment in ASVSpooF 2017 dataset is varying and shown in Table 3.4. To develop an individual environment-dependent replay SSD system, half of the spoof speech utterances for the corresponding environment are chosen for training purpose, and the remaining half are used for testing the performance of the model. To train the genuine speech signal model, equal number of genuine utterances are selected as that of spoof speech utterances, used for training in corresponding environment. To perform experiments for environment-dependent case on ASVSpooF-2019 (PA scenario) dataset, we partitioned the dataset considering the acoustic environment (Table 3.7), and replay configurations (Table 3.8). There are 27 different acoustic configurations, and for each configuration, we have 1070 bonafide utterances and, 7020 spoof utterances. Each half from the bonafide utterances are chosen for training and testing. Among spoof utterances for each acoustic configuration, 2500 utterances are randomly chosen for training and remaining spoof utterances are used for testing. The replay configurations consider the 3 categories of attacker-to-speaker recording distances (D_a), and 3 categories of loudspeaker quality (Q) (defined in Table 3.8). Various combinations of the D_a and Q constitute 9 replay configurations [289]. Only spoofed utterances belongs to either of these configurations. The spoof speech model for each of the replay configuration is trained using half of the utterances belonging to that configuration. The rest half is used for the testing. The genuine speech model is trained using 5400 genuine utterances, and, 23000 genuine utterances are used for testing. The experiments are performed with CQCC and LFCC feature sets.

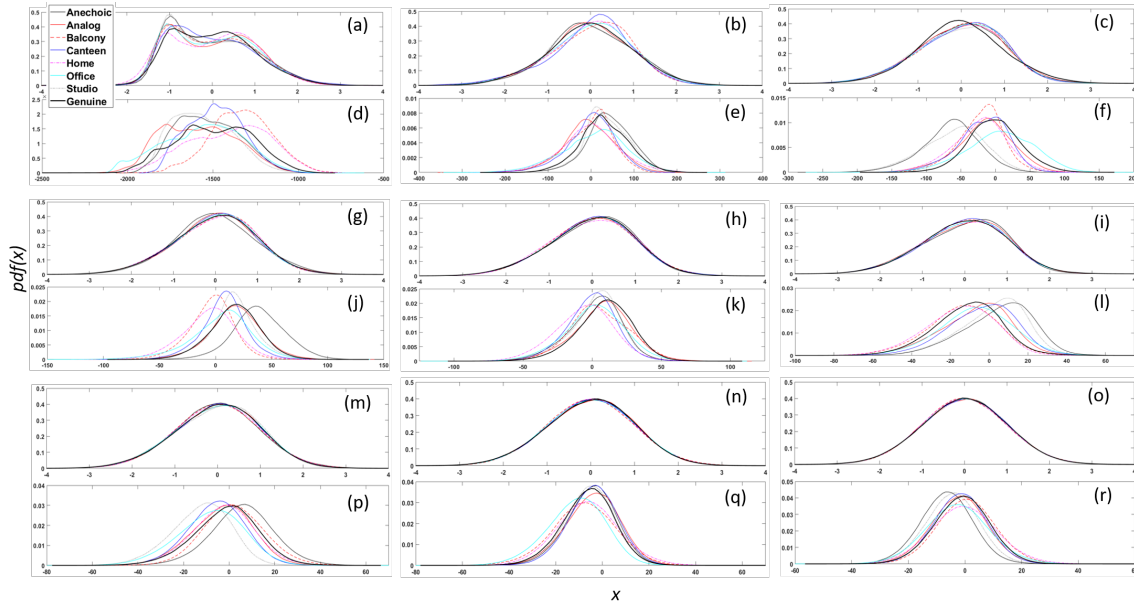


Figure 6.3: Estimated pdf of Genuine and Environmentwise SpooF Speech Samples over the (a) 1^{st} , (b) 3^{rd} , (c) 5^{th} , (g) 10^{th} , (h) 12^{th} , (i) 15^{th} , (m) 20^{th} , (n) 25^{th} , and (o) 30^{th} Feature Dimensions with Application of CMVN, whereas Figure (d), (e), (f), (j), (k), (l), (p), (q), and (r) shows the Estimated pdf s for without CMVN Case with the Same Sequence of Dimensions as that of CMVN Case. Legends of all Figures Are Similar as Given in Figure 6.3(a). After [21].

6.2.5 Experimental Results for ASVSpooF 2017 Dataset

In this Section, we present results to investigate the issue on application of the CMVN technique for SSD task. Performance of the environment-dependent scenario for CQCC and LFCC feature sets using ASVSpooF 2017 dataset is displayed in Table 6.2. The EER is used as the performance evaluation metric. It is observed that, for environment-independent case in ASVSpooF-2017 dataset, CMVN normalization technique works significantly better.

Figure 6.3 shows the estimated pdf for a few selected dimensions of the CQCC feature set for genuine speech samples and spooF speech samples for all the individual environments. Figure 6.3(a), Figure 6.3(b), Figure 6.3(c), Figure 6.3(g), Figure 6.3(h), Figure 6.3(i), Figure 6.3(m), Figure 6.3(n), and Figure 6.3(o) shows the estimated pdf s of the CQCC feature set for 1^{st} , 3^{rd} , 5^{th} , 10^{th} , 12^{th} , 15^{th} , 20^{th} , 25^{th} , and 30^{th} dimensions with CMVN normalization, respectively. Whereas, Figure 6.3(d), (e), (f), (j), (k), (l), (p), (q), and (r) shows the estimated pdf s for 1^{st} , 3^{rd} , 5^{th} , 10^{th} , 12^{th} , 15^{th} , 20^{th} , 25^{th} , and 30^{th} dimensions for without application of the CMVN, respectively. This figure can be used to analyze the behavior of the feature distribution in environment-dependent case. As observed from Figure 6.3, estimated pdf for the CMVN case, all the environments are seems to be

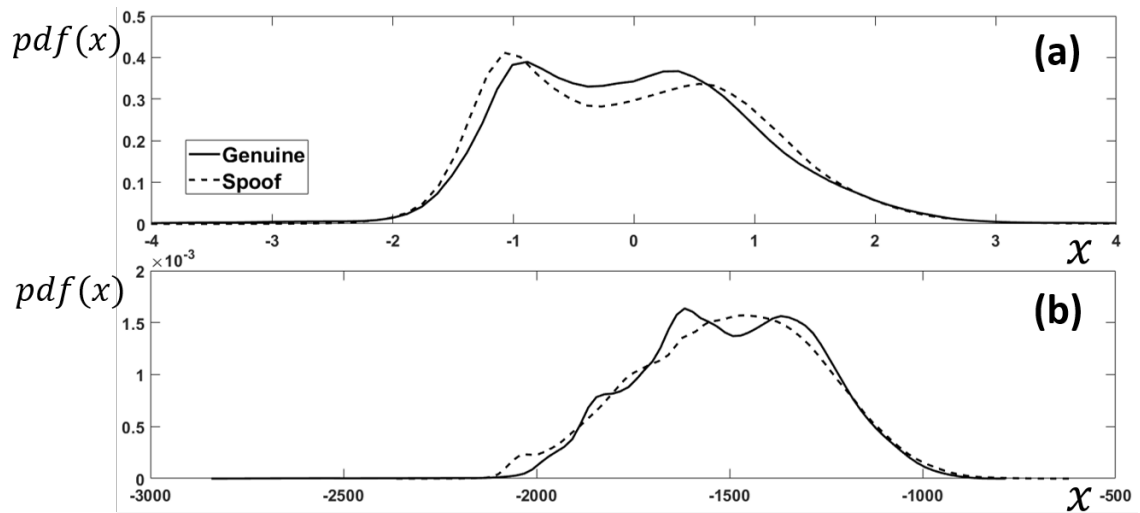


Figure 6.4: Estimated pdf of Genuine and Spoof Speech Samples over the First Feature Dimension for (a) CMVN and (b) without CMVN. Legends of Figure 6.4(b) Are Similar to that of Figure 6.4(a). After [21].

aligned with the pdf of the genuine speech signal. The alignment of the pdf is produced because of CMVN, which further leads to degradation of the results for environment-dependent scenario as distribution of the genuine speech data seems to be similar to that of the spoof speech data. Observations are also made for other dimensions of the feature vector other than mentioned dimensions, however, $pdfs$ for a few selected dimensions are presented in Figure 6.3. It can be observed that, after 11th dimension $pdfs$ of the spoof speech data for the CMVN case, all the other environments mostly aligned with the genuine data and almost no difference exists in their $pdfs$. This fact can be observed from Figure 6.3(k), (l), (p), (q), and (r), where all the $pdfs$ are almost merged. Whereas, the $pdfs$ of genuine data is much different to that of individual spoof speech environments in without CMVN case. Furthermore, for without normalization scenario, the distinct difference in $pdfs$ can be observed for almost all the dimensions. For without CMVN case, if GMM parameters (i.e., *mean* and *variance* of the Gaussian mixtures) are estimated for the $pdfs$ of the genuine *vs.* any other environment for spoof speech data, then for most of the environments, we obtained the well distinguishable GMM parameters. In particular, it can be observed that GMM parameters of the genuine data *vs.* spoof speech data from balcony/studio would be well distinguishable. Hence, corresponding SSD systems are producing 0 % EER. However, with CMVN applied to a feature set, all the $pdfs$ of spoof speech signal representations for an individual environment, do not lie on either side of the pdf of genuine signal representations. This fact is also observed from Figure 6.2(a), where the data samples for genuine speech signal lie in the middle of the

Table 6.1: Results of CQCC-GMM and LFCC-GMM Systems in % EER for Environment-Independent Case on ASVSpooof 2017 Dataset. After [21].

		Dev	Eval
CQCC	Without CMVN	10.31	28.02
	CMVN	12.48	18.17
LFCC	Without CMVN	7.02	32.62
	CMVN	14.79	14.83

other two spoof speech environments. Hence, cumulative distribution of all the environments (shown in Figure 6.4) for spoof speech data, could not produce the distinguishable GMM parameters w.r.t. genuine data.

Figure 6.4(a) and (b) shows estimated pdf of the genuine *vs.* spoof speech signal over the first dimension of the CQCC feature set, obtained by application of the CMVN and without CMVN, respectively. Here, spoof speech data is obtained from all the possible environments. In this case, if GMM parameters are estimated from the $pdfs$, then Figure 6.4(a) will have the more distinguishable GMM parameters as the GMM parameters estimated from this pdf would be better separated than the case of without CMVN. It is because of the fact that, pdf maxima of this pdf are well separated in Figure 6.4(b), many local maxima are observed and random variable corresponding to these maxima are mixing with each other. Because of these closely-spaced GMM parameters for genuine and spoof speech signals, the classifier model may pose ambiguity, when a test sample is presented to the trained model for the SSD task. This might be the reason for getting better results for environment-independent case with application of the CMVN compared to without CMVN case. Authors believe that pdf corresponding to higher cepstral dimensions show less discrimination between genuine *vs.* spoof due to decay of cepstrum w.r.t. time [265,270,271].

Figure 6.5(a) and Figure 6.5(b) shows the DET curves for the system developed using the feature set with and without application of the CMVN on ASVSpooof 2017 challenge dataset, respectively [179]. These DET curves are shown for environment-dependent scenario. Two systems are showing 0 % EER, which cannot be observed on the DET curve. However, these plots are shown by the point at the origin. It can be observed from the DET curves that the performance of environment-dependent case is significantly improved without normalization of the feature set.

Table 6.2: Results of CQCC-GMM System in % EER for Environment-Dependent Case on ASVSpooF 2017 Dataset. After [21].

	CQCC		LFCC	
	CMVN	Without CMVN	CMVN	Without CMVN
Anechoic Room	10.02	0.26	10.60	0
Analog wire	16.99	11.42	22.09	10.89
Balcony	13.81	0	9.60	0.13
Canteen	3.43	0.93	2.73	1.33
Home	7.39	2.12	9.23	2.51
Office	14.99	5.63	17.62	7.22
Studio	7.53	0	7.21	0

6.2.6 Experimental Results for ASVSpooF 2019 dataset

Figure 6.6 shows the *pdfs* estimated from the first cepstral (trajectory) coefficient of the CQCC feature set for the genuine, replay spooF, and individual replay configurations. As the total number of utterances for the genuine and each replay configuration is large, we randomly selected 5000 utterances of the genuine and individual replay configurations, to estimate the *pdfs* shown in Figure 6.6. The analysis is performed for various dimensions, however, the first cepstral trajectory is selected as the sample example, as observations are clearer for the first dimension than the other dimensions. Figure 6.6(a) and Figure 6.6(c) shows the estimated *pdfs* for the genuine *vs.* replay speech samples without and with applying the CMVN, respectively. Further, the results obtained on standard dataset and protocols is shown in Table 6.3. From Figure 6.6(a), it can be observed that the *pdfs* of bonafide and spooF is more aligned with each other than that of Figure 6.6(c). It shows that the feature set without CMVN shows the better classification capability over the normalized feature set. These observations are reflected via the performance of the SSD system as shown in Table 6.3, where the CMVN shows the degraded performance than the without normalized feature set. Figure 6.6(b) shows estimated *pdfs* of CMVN for the bonafide *vs.* individual replay configurations, where it can be observed that all of the individual *pdfs* of the replay configurations are aligned with each other and also with *pdf* of the bonafide feature set. Hence, the degraded performance is observed for the environment-dependent SSD task with application of the CMVN, as shown in Table 6.4. Figure 6.6(d), Figure 6.6(e), and Figure 6.6(f) shows the *pdfs* estimated for without normalized feature sets of the bonafide *vs.* individual replay configurations. It can be

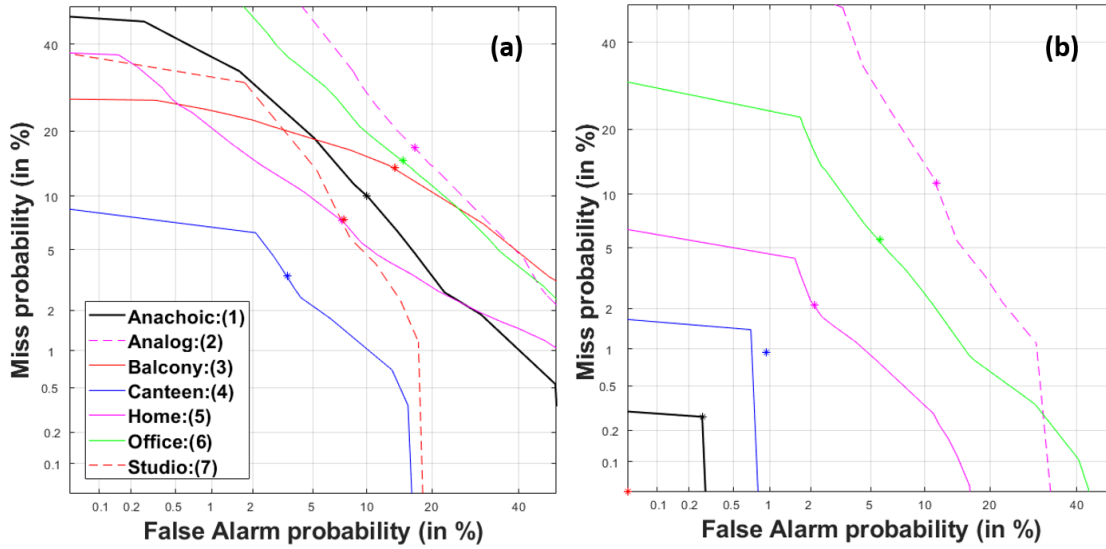


Figure 6.5: DET Plots for Environment-Dependent Case using ASVSpooF-2017 Dataset (a) with Application of the CMVN, and (b) without Application of the CMVN on Feature Set. Legends for Figure 6.5(a) and Figure 6.5(b) Are the Same. After [21].

observed that for perfect replay device quality (Q), the *pdfs* of the replay configuration features are very much aligned to that of the *pdf* of the bonafide features. As Q degrades, the difference between the *pdfs* of the bonafide *vs.* replay configurations, is more vivid. This variation in the distribution characteristics is reflected into the performance of the replay SSD task for environment-dependent scenario, as shown in Table 6.4. It is also clear from Table 6.4 that the application of the CMVN for environment-dependent case significantly degrades the performance of the SSD system.

Table 6.3: Results of CQCC-GMM Systems ASVSpooF 2019 Dataset using Standard Protocols.

		Dev	Eval
ASVSpooF-2019	without CMVN	9.48	11.58
	CMVN	12.42	13.78

From Table 6.3, it can be observed that application of the CMVN significantly improves the performance on ASVSpooF-2017 challenge dataset, whereas it degrades the performance on ASVSpooF-2019 challenge dataset. This mysterious results can be unfolded by comparing the *pdfs* shown in Figure 6.6 with *pdfs* estimated for ASVSpooF-2017 challenge dataset as reported in [21]. It can be clearly observed that for ASVSpooF-2017 challenge dataset that the *pdfs* of the bonafide data is much different to that of the individual spooF speech environments for

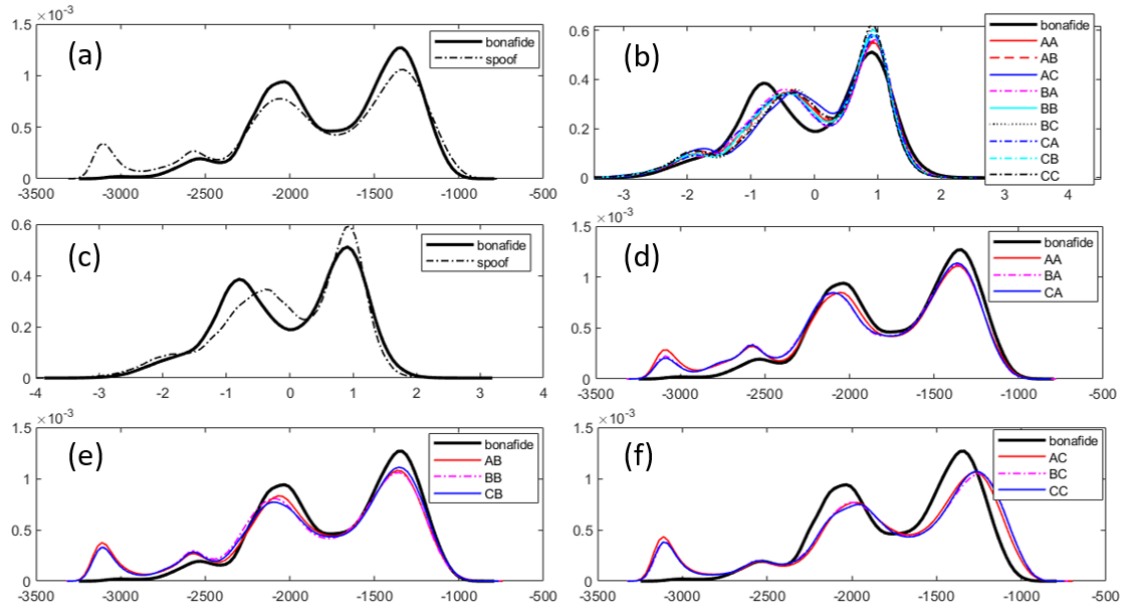


Figure 6.6: All the $pdfs$ Shown in Figure 6.6 Are Estimated from 1^{st} Cepstral Coefficient of the CQCC Feature Set. Figure 6.6(a) and Figure 6.6(c) Shows the Estimated $pdfs$ for the Genuine *vs.* Spoof Speech Samples without Normalization, and CMVN, Respectively. Figure 6.6(b) Shows the Estimated $pdfs$ for the Genuine *vs.* Individual Replay Configurations with CMVN Applied on CQCC Feature Set. Whereas, Figure 6.6(d), Figure 6.6(e), and Figure 6.6(f) Shows the Estimated $pdfs$ for the Genuine *vs.* Individual Replay Configurations without Normalization Applied on CQCC Features. After [21].

without CMVN case. Hence, it produces better performance for environment-dependent scenario. However, the $pdfs$ for individual spoof speech environments lie on both the sides of the bonafide pdf . Hence, cumulative effect of all individual environments is aligned with bonafide pdf , which results in degradation in the results for without CMVN case. This is not the case here for ASVSpooF-2019 dataset. It can be observed from Figure 6.6 that for without CMVN case, the $pdfs$ of the individual replay configurations are lying on the one of the side to that of bonafide pdf . In addition, as the Q degrades, $pdfs$ of the bonafide data and individual replay configurations are separated away from each other. This might be due to the fact that simulated replay mechanism systematically alter the replay configuration characteristics, which results in shifting the similar $pdfs$ in a particular direction, as shown in Figure 6.6(d), Figure 6.6(e), and Figure 6.6(f). Hence, their cumulative effect will not be as undesirable as that of ASVSpooF-2017 case. This is an important finding of its kind as it can easily predict whether or not one should use the CMVN for classification task. This finding of ours is in agreement with the fact that CMVN acts like a *double edged sword* as observed originally in the speaker

Table 6.4: Results (in % EER) for Environment-Dependent Case for Various Replay Configurations on ASVSpooF 2019 Challenge Dataset.

Replay configuration	Without CMVN	CMVN	Replay configuration	Without CMVN	CMVN
AA	23.81	26.94	BA	22.54	30.86
AB	2.01	10.60	BB	2.01	10.35
AC	0.85	4.73	BC	0.75	5.17
CA	21.83	28.85	CB	1.77	11.19
CC	0.89	5.09	-	-	-

recognition literature [281,282]. In particular, such feature normalization via CMN perform better if training and testing is done with speech recordings in different transmission channel, acoustic environments, etc., whereas CMN could hurt the classification/recognition performance if training and testing is done with speech recordings in the same acoustic environments and hence, the same transmission channel characteristics. Thus, such feature normalization should be applied very carefully considering the type of noise, recording conditions, and application at hand.

The SSD experiments are also performed for environment-dependent case for the acoustic configurations, given in Table 3.7. From Table 6.5, it can be observed that the SSD system with CMVN performs better than without CMVN case. This trend is as similar as observed for the replay configurations case as observed in Table 6.4.

From the results obtained in this study on ASVSpooF 2017 and ASVSpooF 2019 datasets, we observed that the application of the CMVN to the feature set do not always guarantee better results for the classification task. However, it depends upon the variability of the speech samples in terms of transmission channel noise. The contradictory behavior of CMVN is observed on these two datasets. For ASVSpooF-2019 dataset, it can be observed that the *pdfs* of the individual replay configurations are lying on one of the sides than that of bonafide *pdf* for without CMVN case. Furthermore, the *pdfs* of the replay configurations are separated apart from the *pdf* of the bonafide data, as the replay configuration characteristics are intensified. By applying CMVN, these replay characteristics are suppressed and bringing the *pdfs* of bonafide and spoof data closer to each other, which loses the classification capability to a certain extent and thus, degradation in the performance of SSD systems. This might be because of the generation of the replayed speech samples by simulating the acoustic and replay configurations, rather than real replay environments in ASVSpooF 2017 counterpart. However, in ASVSpooF-

Table 6.5: Results in (% EER) for Environment-Dependent Case for Various Acoustic Configurations on ASVSpooF 2019 Challenge Dataset.

Replay configuration	Without CMVN	CMVN	Replay configuration	Without CMVN	CMVN
aaa	10.16	16.34	bbc	3.57	5.69
aab	7.707	15.04	bca	3.04	3.28
aac	5.73	15.36	bcb	2.07	2.58
aba	3.71	3.78	bcc	1.39	1.79
abb	3.45	4.37	caa	13.92	19.58
abc	3.67	5.53	cab	13.93	21.83
aca	1.67	1.98	cac	11.84	20.33
acb	1.23	1.97	cba	4.78	5.16
acc	0.76	1.17	cbb	6.13	6.80
baa	11.11	16.57	cbc	5.59	6.71
bab	11.13	18.57	cca	3.90	3.12
bac	9.10	17.71	ccb	2.08	2.80
bba	3.67	2.90	ccc	2.61	2.56
bbb	5.11	4.30	-	-	-

2017 challenge dataset for without CMVN case, *pdfs* of the individual replay environments are lying on both the sides of the *pdf* of the bonafide data. Hence, the cumulative effect of all the replay environments might create difficulty for the classification of bonafide *vs.* spoof speech. However, the results are similar for the application of CMVN in environment-dependent cases for both the datasets. Finally, we conclude that the applicability of the CMVN on cepstral features for the SSD task depends upon the intended dataset, which can be analyzed using the *pdfs* of the sample data.

6.3 DAS *vs.* MVDR Beamformer: Analysis for Replay SSD Task

6.3.1 Motivation

As we studied in Chapter 5 (Section 5.3.1), the replay mechanism consists of the characteristics of the recording, playback devices, and corresponding acoustic environments due to which reverberation characteristics are embedded into the replay spoof signal [72]. This study investigate the capability of the DAS *vs.* MVDR beamformer to extract the reverberation characteristics in replay speech signals,

which can be utilized for replay SSD task [41, 106, 290, 291]. MVDR is a state-of-the-art beamformer for speech enhancement applications as it successfully *nullify* the reverberation effects in distant speech signals [290, 291]. Whereas, DAS suppresses the additive noise and retains the reverberation effect observed in the output signal [292] and hence, we deduce that DAS is a suitable choice for replay SSD task. This hypothesis is validated using experimental results on ReMASC. Furthermore, TECC feature set being capable of capturing the reverberation effects, is used effectively in *tandem* with DAS beamformer for replay SSD task [41].

6.3.2 Signal Modeling for Microphone Array Signal

Assuming the LTI model for the acoustic medium (path) between speech sound source and microphone array, the speech signal received by N -element microphone array is given as [42, 270, 290, 293]:

$$\begin{aligned} x_i(n) &= r_i(n) * k(n) + \eta_i(n), \\ &= y_i(n) + \eta_i(n), i = 1, 2, \dots, N, \end{aligned} \quad (6.12)$$

where i represents the index for i^{th} microphone in an array, $r_i(n)$ is the impulse response of the acoustic medium between the desired source signal $k(n)$ and i^{th} microphone. $*$ represents the convolution operation and $\eta_i(n)$ corresponds to additive noise of the i^{th} microphone. Here, for modeling of noisy speech signal $x_i(n)$, it is assumed that the speech signal $y_i(n)$ and noise signal $\eta_i(n)$ are zero-mean and uncorrelated. During development of the replay speech signal, impulse responses of recording devices ($rd(n)$), and environment ($re(n)$) as well as impulse responses of playback devices ($pd(n)$), and environment ($pe(n)$) are convolved with the source signal. Let $N(n)$ represents the combination of these impulse responses [72], i.e.,

$$N(n) = rd(n) * re(n) * pd(n) * pe(n). \quad (6.13)$$

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

$$\begin{aligned} x_{ir}(n) &= r_i(n) * N(n) * k(n) + \eta_i(n), \\ &= y_{ir}(n) + \eta_i(n), i = 1, 2, \dots, N. \end{aligned} \quad (6.14)$$

Thus, the characteristics of the $y_{ir}(n)$ in eq. (6.14) is different from that of $y_i(n)$ because of the additional impulse response $N(n)$ caused by the replay mechanism. Considering this $N(n)$ as distinguishing acoustic characteristics of the re-

play spoof, it can be emphasized using suitable signal processing technique for replay SSD. To that effect, first we present the significance of the DAS beamformer over MVDR for replay SSD through mathematical analysis, and then it is validated using experiments.

The representation of the received signal in eq. (6.12) in the frequency-domain can be expressed as [293]:

$$\begin{aligned} X_i(\omega) &= R_i(\omega) \odot K(\omega) + H_i(\omega), \\ &= Y_i(\omega) + H_i(\omega), \quad i = 1, 2, \dots, N, \end{aligned} \quad (6.15)$$

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

$$\mathbf{X}(\omega) = \mathbf{R}(\omega) \odot \mathbf{K}(\omega) + \mathbf{H}(\omega) = \mathbf{Y}(\omega) + \mathbf{H}(\omega), \quad (6.16)$$

where

$$\begin{aligned} \mathbf{X}(\omega) &= [X_1(\omega), \dots, X_N(\omega)]^T, \mathbf{R}(\omega) = [R_1(\omega), \dots, R_N(\omega)]^T, \\ \mathbf{K}(\omega) &= [K(\omega), \dots, K(\omega)]^T, \mathbf{Y}(\omega) = [Y_1(\omega), \dots, Y_N(\omega)]^T, \\ \mathbf{H}(\omega) &= [H_1(\omega), \dots, H_N(\omega)]^T. \end{aligned} \quad (6.17)$$

6.3.3 Delay and Sum (DAS) Beamformer

The DAS is a primitive beamforming technique for noise reduction in the array signal processing literature [292, 294]. This involves reinforcing the desired signal while suppressing the unwanted noise signals. The conventional DAS beamformer will delay all the input signals in time *w.r.t.* the reference signal, such that the array sensor can focus in one direction. Hence, the summation of the delayed signals with the reference signal will result in suppression of noise, which is arriving from the other directions. Furthermore, it can be postulated that the summation of the delayed signals leads to cancellation of *additive (random) noise*. Figure 6.7 shows the functional block diagram of DAS beamformer from receiver end. Here, weights for corresponding single channel microphone signal in a microphone array are shown. The time-domain representation of DAS beamformer is given by [295]:

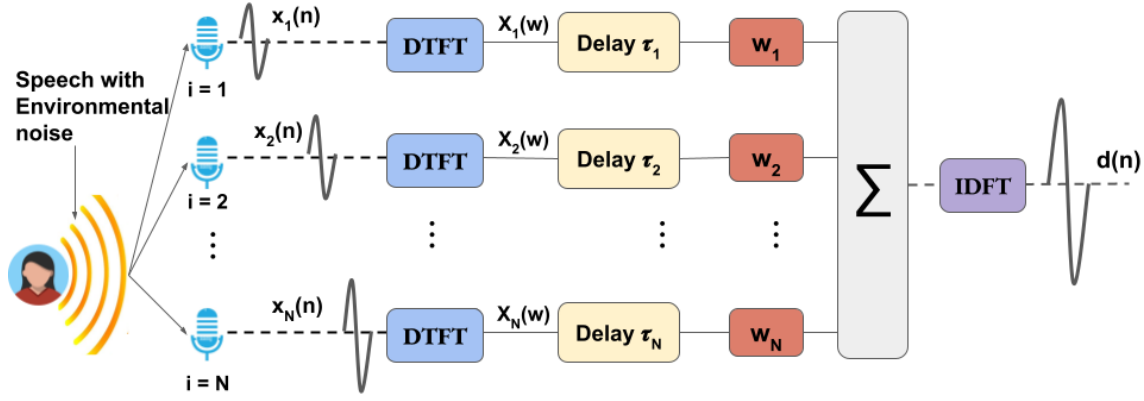


Figure 6.7: Functional Block Diagram of DAS Beamformer having N Number of Microphones in an Array. After [37].

$$\mathbf{d}(n) = \frac{1}{\beta} \sum_{i=1}^N w_i x_i(n - \tau_i). \quad (6.18)$$

Furthermore, the frequency-domain representation of DAS beamformer is given by taking DTFT of eq. (6.18) [296]. In particular,

$$\mathbf{D}(\omega) = \frac{1}{\beta} \sum_{i=1}^N w_i X_i(\omega) e^{-j\omega\tau_i} = \mathbf{W}^H \mathbf{X}(\omega), \quad (6.19)$$

$$\text{where } \mathbf{W} = \frac{1}{\beta} \sum_{i=1}^N w_i e^{-j\omega\tau_i}, \quad (6.20)$$

where w_i is the elementwise weighting for the spatial window, β is the summation of the weights, and \mathbf{W} is the steering vector (optimized weight vector) of desired *linear* phase shift and weights. The superscript H denote the Hermitian transpose. In fact, it should be noted that it is due to this linear phase filtering, acoustic characteristics of replay are preserved in DAS beamformed signal. The $\mathbf{D}(\omega)$ represents the frequency response of the beamformed signal. In the framework of Wiener-Khinchin theorem [249], the power at the output of the beamformer is estimated by taking the Fourier transform of autocorrelation function of the beamformer output [249], i.e.,

$$\mathbf{p}(\omega) = E[|\mathbf{D}(\omega)|^2], \quad (6.21)$$

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

6.3.4 Minimum Variance Distortionless Response (MVDR)

MVDR beamformer achieves the speech enhancement by suppressing (ideally nullifying) the reverberation effects introduced by the room acoustics [290, 291]. In this approach, Signal-to-Noise Ratio (SNR) of the multi-channel audio signal is significantly improved by minimizing the distortion (noise) [297]. For this formulation, it is assumed that the audio signal from the reference source is distortionless, which also results in preservation of all-pass characteristics. However, MVDR increases the computational complexity of the system. The matrix for output power $\mathbf{p}(\omega)$ of MVDR beamformer is given by:

$$\mathbf{p}(\omega) = E[|\mathbf{D}(\omega)|^2] = \mathbf{W}^H \mathbf{V}(\omega) \mathbf{W}, \quad (6.22)$$

where $\mathbf{V}(\omega)$ and \mathbf{W} represents the matrix of cross-power spectral density and initial weight matrix, respectively. The co-variance matrix for L number of frames is given by [295]:

$$\hat{\mathbf{V}}(\omega) = \frac{1}{L} \sum_{l=0}^{L-1} \mathbf{x}_l(\omega) \mathbf{x}_l^H(\omega), \quad (6.23)$$

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

$$\begin{aligned} \underset{\mathbf{W}}{\operatorname{argmin}} \quad & \mathbf{W}^H(\omega) \hat{\mathbf{V}}(\omega) \mathbf{W}(\omega), \\ \text{subject to} \quad & \mathbf{W}^H(\omega) \mathbf{m} = 1, \end{aligned} \quad (6.24)$$

where \mathbf{m} represents the steering vector, which is the most crucial matrix for direction estimation of the desired signal. It provides the directional information of microphone array. During this minimization, it affects the impulse response of the underlying acoustic medium. Let d_i be the desired direction representation for the element i . Then, the steering vector for i^{th} element (i.e., m_i) is given by [296]:

$$m_i = e^{j\omega d_i}. \quad (6.25)$$

Constrained minimization in eq. (6.24) is performed by using Lagrange multipliers [298]. Hence, the optimum weight matrix for MVDR beamformer is given by [296]:

$$\mathbf{W}_o(\omega) = \frac{\hat{\mathbf{V}}^{-1}(\omega) \mathbf{m}}{\mathbf{m}^H \hat{\mathbf{V}}^{-1}(\omega) \mathbf{m}}. \quad (6.26)$$

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

$$\mathbf{D}(\omega) = \mathbf{W}_o^H(\omega)\mathbf{X}(\omega). \quad (6.27)$$

Furthermore, the output power ($\mathbf{p}_o(\omega)$) of MVDR beamformer is given by [296]:

$$\mathbf{p}_o(\omega) = \mathbf{W}_o^H(\omega)\hat{\mathbf{V}}(\omega)\mathbf{W}_o(\omega). \quad (6.28)$$

6.3.5 Reverberation Analysis Using TEO

In this study, we have analyzed the effect of reverberation in genuine *vs.* replayed speech signals via its time-domain representation and TEO profile. The study reported in [41] shows that the TEO profile effectively captures the characteristics of reverberation in the replay signal (which includes different reflections of the genuine signal [299, 300]), introduced during its recording from the far-field. In particular, there are multi-order reflections possible depending on the recording environment [41]. In Figure 6.8, the time-domain representation of genuine (Figure 6.8(c)) and replay (Figure 6.8(d)) signals are shown. The Figure 6.8(a) and Figure 6.8(b) represents the zoomed version of the dotted squared region from Figure 6.8(c) and Figure 6.8(d), respectively. Furthermore, Figure 6.8(e) and Figure 6.8(f) corresponds to the zoomed version of the solid squared region from Figure 6.8(c) and Figure 6.8(d), respectively. Hence, from this zoomed figures, it can be observed that the replayed signal has additional impulses and distortions as compared to the genuine speech, which are due to the added reverberation [41, 106].

Furthermore, in Figure 6.9, we show the TEO profiles in order to observe the Teager energy traces for genuine (Panel I) *vs.* replay (Panel II) for (a) ReMASC and its (b) DAS *vs.* (c) MVDR beamformed versions. The additional energy traces can be observed in Figure 6.9 for DAS over original ReMASC and its MVDR beamformed version via the oval and square boxes for both genuine and replay signals, respectively. This observation is further validated via experiments performed on TECC feature set for original ReMASC and its DAS *vs.* MVDR beamformed versions. The results in Table 6.6 shows that TECC-GMM system applied on DAS beamformed signals gives relatively best performance over all the other combinations.

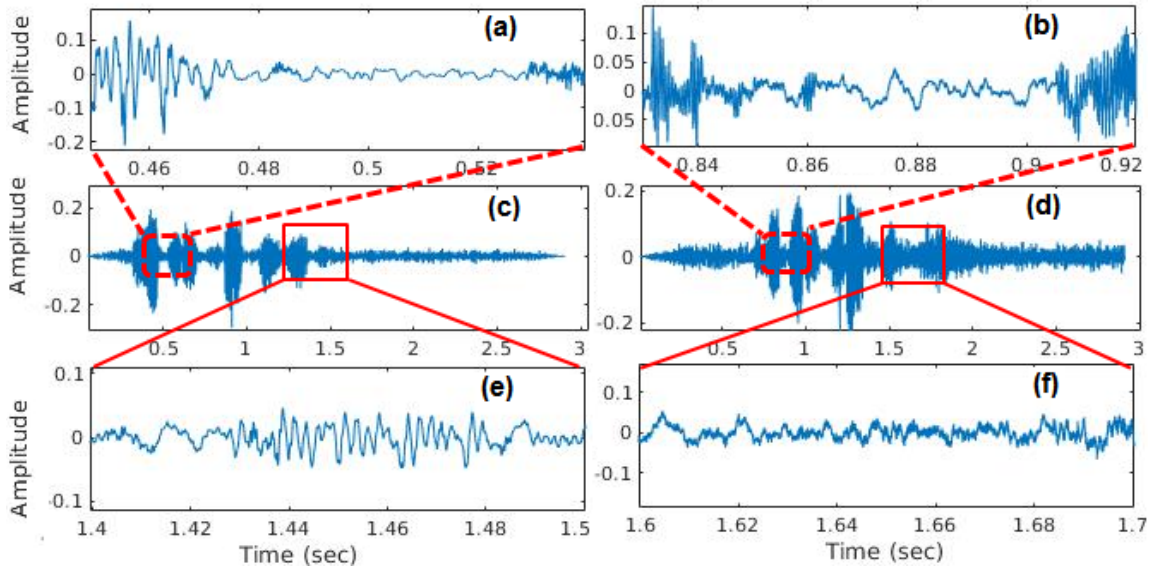


Figure 6.8: Time-domain Representation of (c) Genuine *vs.* (d) Replayed Speech Signal from ReMASC Dataset. Figure 6.8(a) and Figure 6.8(b) Represents the Zoomed Version of the Dotted Squared Region and Figure 6.8(e) and Figure 6.8(f) Corresponds to the Zoomed Version of the Solid Squared Region from Figure 6.8(c) and Figure 6.8(d), Respectively. After [22].

6.3.6 Experimental Setup

For this task, we utilized the ReMASC dataset with data distribution as shown in Table 3.15. The similar configuration was utilized in [11, 301]. The brief details of the dataset are discussed in Chapter 3. The experiments are performed using single channel microphone signal in microphone array and the two beamforming techniques, i.e., DAS and MVDR. The state-of-the-art CQCC, MFCC, LFCC along with TECC feature sets are utilized in this study. The GMM, CNN, and LCNN classifiers are employed in this study. The brief details of the feature sets and classifiers can be studied from the Chapter 3.

6.3.7 Experimental Results

The performance of DAS *vs.* MVDR beamformer is evaluated using % EER. The SSD systems are developed for CQCC, LFCC, and TECC feature sets using GMM, CNN, and LCNN-based classifiers for all the three datasets, i.e., ReMASC and its DAS *vs.* MVDR beamformed versions. The % EER on development and evaluation sets are shown in Table 6.6 for all the three variants of datasets. It was observed that *only static* features performed better than all the other combinations for this dataset. Hence, all the results reported in Table 6.6 are obtained using only static features. Furthermore, improved performance is obtained on the DAS

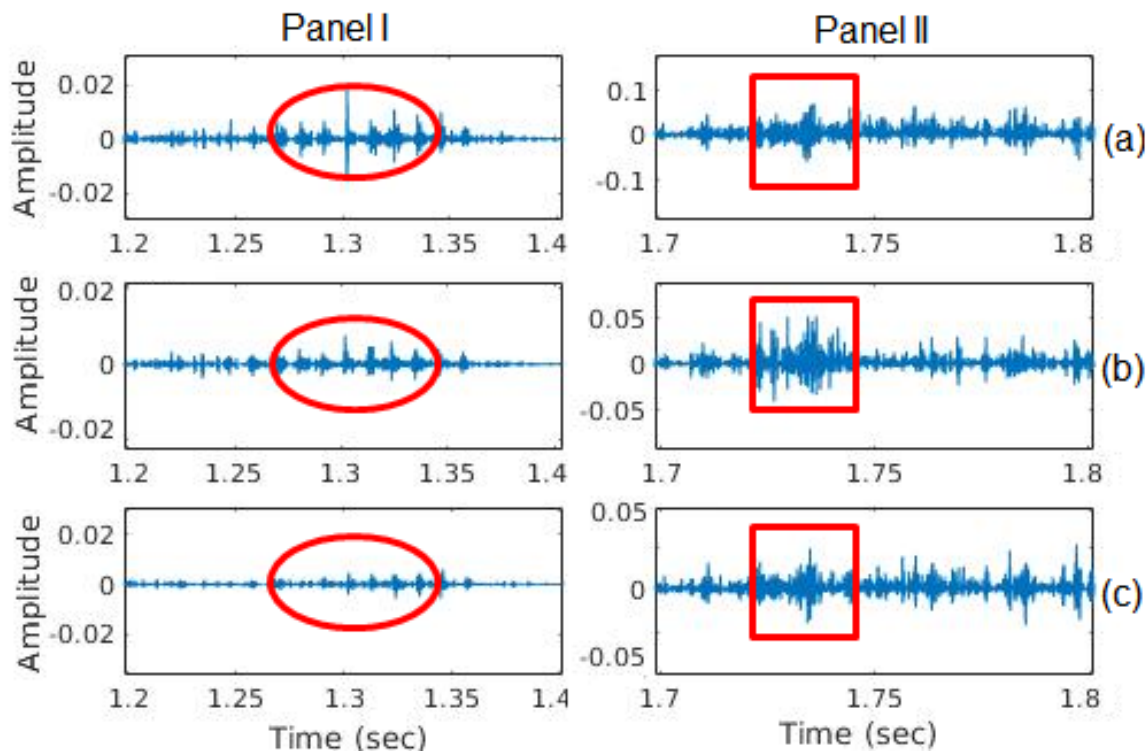


Figure 6.9: TEO Profile of Genuine (Panel I) and Replayed (Panel II) Speech Signals from (a) Original ReMASC and its (b) DAS *vs.* (c) MVDR Beamformed Versions. After [22].

beamformed dataset than that for the original ReMASC and MVDR beamformed version, for all the feature sets and classifiers considered in this study. This suggests that the DAS beamforming can be potentially utilized to improve the performance of the replay SSD system for VAs. In addition, the TECC feature set performs better than that of other feature sets for all the classifiers and all the dataset versions. This proves the capability of TECC to extract the reverberation characteristics in replay speech signal. In particular, relatively the best performance is observed for TECC-GMM SSD system for DAS beamformed dataset. It should also be noted that, results of MVDR are worse than unprocessed (i.e., not beamformed) ReMASC data. Furthermore, the performance of all the systems are also shown using DET curves in Figure 6.10. In particular, Figure 6.10(a) and Figure 6.10(b) shows the DET curves for Dev and Eval sets, respectively, for TECC-GMM system on all the three versions of datasets. It can be observed from Figure 6.10 that the DAS beamformed dataset consistently performing well as compared to the original ReMASC and its MVDR beamformed version for both Dev and Eval sets.

Thus, from our deduction and experimental results, it can be observed that this work is contradictory w.r.t. suitability of state-of-the-art MVDR beamformer for

Table 6.6: Results (in % EER) on ReMASC and its DAS *vs.* MVDR Beamformed Versions using Various Feature Sets and Classifiers. After [22].

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

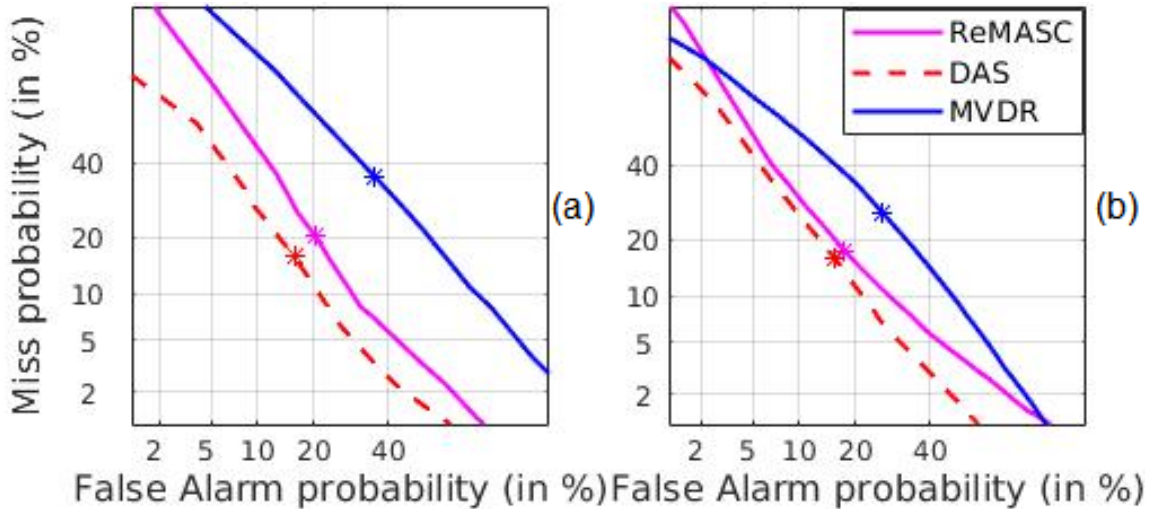


Figure 6.10: DET Curves for ReMASC and its Beamformed Versions using TECC with GMM: (a) Dev, and (b) Eval set. After [22].

Distant Speech Recognition (DSR), indicating a straightforward generalization of beamforming method from DSR to replay SSD in VAs is *not* recommended even though DSR is very much integral part of VAs. In addition, due to linear phase characteristics of DAS beamformer, the acoustical characteristics of reverberation in replay spoof are presented and hence, TECC is employed to capture these reverberation characteristics.

6.4 Severity-Level Classification of Dysarthric Speech

6.4.1 Motivation

Dysarthria [302] consists of a motor speech disorder in which the articulatory elements and muscles required to speak ordinarily are somehow affected, paralyzed or damaged. Individuals suffering from dysarthria face difficulties in conveying a spoken message or expressing voice emotions, since the vocal folds, tongue, and associated muscles cannot be adequately controlled. Furthermore, most of the VAs are developed based on the assumption that speech signal comes from the healthy subjects [303, 304]. Hence, they may be suitable to develop speech technology applications efficiently on impaired or disabled people [305–307].

To allow for dysarthric speech enhancement and patients' progression in treatment, detecting the severity-level of a pathology from a short-duration speech segments is an essential task. Standard methods for the assessment of dysarthric speech are traditionally based on a clinical trial by speech language pathologists, using pre-defined rating scales or observing the movement of various articulatory elements over the spoken time interval [308]. For dysarthria detection, specific speech segments from a certain set of speakers can be obtained from longer duration speech signals based on manual or automatic framing. In the latter case, deep learning-based approaches have played an important role, as demonstrated in papers [309], [310], and [311]. For the severity-based classification, most of the methods focuses on feature-based techniques and acoustic modeling [312–321]. One of the conventional classifier-based approaches can be studied in [315, 322]. However, deep learning-based approaches can be studied in [307, 323–329]. In addition, inspired by two recently-proposed CNN-based approaches used to detect Parkinson's disease, as documented in papers [39] and [329], our strategy focus on a ResNet-based classification algorithm.

6.4.2 Problem Formulation

Our goal is to classify dysarthric speech based on its severity-level using short-duration speech segment. In this study, dysarthric speech is classified in four severity-level-based categories as shown in Table 6.7. As suggested in the [23], five naive listeners were recruited for each speaker, and they were allowed to listen to words as many times as needed for transcription. For each listener's transcription, the percentage of correct responses was calculated. The correct percentage was then averaged across five listeners to obtain each speaker's intelligibility. Based

on the averaged percent accuracy, each speaker was classified into one of four categories as shown in Table 6.7.

Table 6.7: Severity-Level Classification Based on Intelligibility. Adapted from [23].

Intelligibility Rating (%)	Severity-Level
0-25	High
25-50	Medium
50-75	Low
75-100	Very Low

Since speech is essentially produced during air exhalation, a precisely coordinated respiratory support is of paramount importance for communication. In dysarthric subjects, however, the combined pneumo-phono-articulatory cognitive commands from the brain are pathologically-affected, drastically degrading speech quality. Remarkably observable, distorted vowel sounds have been a direct consequence of dysarthria, where articulatory undershoot forces a humble vowel working space, as mentioned in papers [330] and [331]. Hence, as shown in papers [332], [333], [334], [335], and [336], formant frequencies centralization, uncommon formant frequencies for both front and high vowels, formants instability, and reduced slopes involving the second formant, are notable. This justifies our efforts in using short speech segments for the detection of dysarthria, since, presumably, they contain the formant-related information we need and, in addition, are capable of characterizing severity-levels. Based on our strong evidences, let us move forward to the formal problem formulation.

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ denotes the features of dysarthric speech, and $\mathcal{Y} = \{y_i\}_{i=1}^n$ denotes the corresponding labels. First, we map this labeled data, $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ as:

$$f(x) = \begin{cases} y = 0, & \text{if severity-level is high,} \\ y = 1, & \text{if severity-level is mid,} \\ y = 2, & \text{if severity-level is low,} \\ y = 3, & \text{if severity-level is very low.} \end{cases}$$

Problem Statement: Given a manually annotated dysarthric speech data (\mathcal{D}), for severity-based classification in four categories, learn severity-based classifier (as a mapping function), $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$, which can do efficient classification using short-duration speech.

To solve our proposed problem, we certainly need an adequate classifier. Although universal approximation theory [337] presents results allowing for the

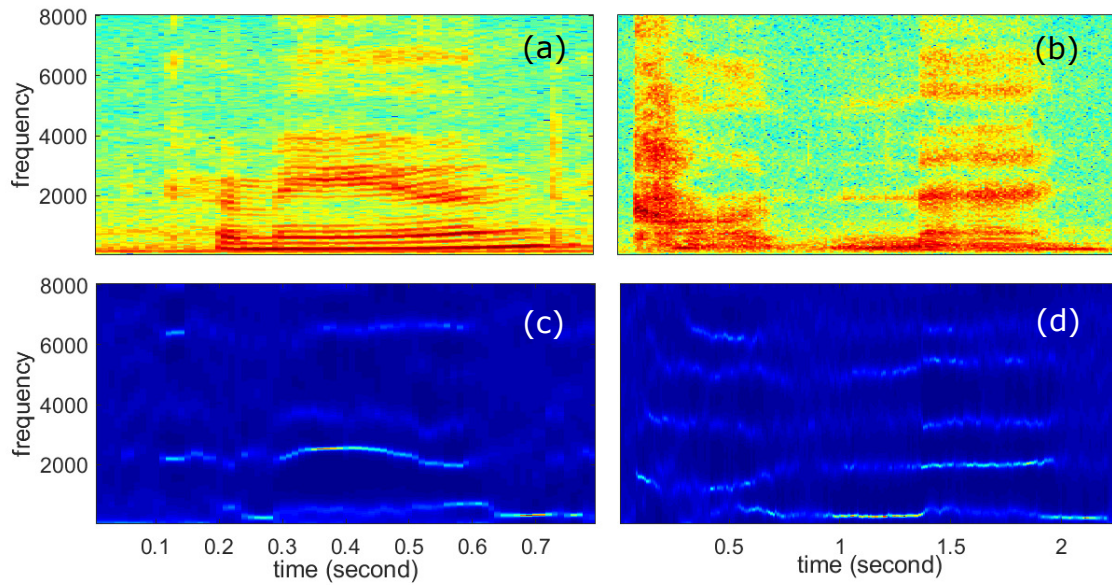


Figure 6.11: STFT Representation of: (a) Normal Speech, (b) Dysarthric Speech *vs.* LP Spectrum of (c) Normal Speech, and (d) Dysarthric Speech. After [24].

conclusion that feedforward neural networks containing a single layer could represent any function, data overfitting and the vanishing gradient issues have forced machine learning algorithms to advance much more. As observed in practice and confirmed theoretically, however, expanding the networks in such a way they get deeper does not mean just adding layers because accuracy and performance might degrade extraordinarily fast. Thus, since they allow for training up to thousands of layers with remarkable performance, ResNets [338] have been considered one of the most groundbreaking advancements in deep neural network-related fields. ResNet was briefly discussed in Chapter 3 and used for anti-spoofing research in this thesis (Chapter 5, Section 6.5.4.2).

6.4.3 Characterizing Dysarthria in Speech Signals

Features derived from time-frequency representation of speech signal have been used in several speech applications. In particular, the study in [339] evaluated various acoustic features based on their relative effectiveness to estimate quality of time-frequency mask, namely, ideal binary mask (IBM) - a central research issue in speech enhancement and source separation area. The wide range of acoustic features (primarily motivated from robust ASR), such as MFCC, PLP, RASTA-PLP, GFCC, PNCC, F_0 , etc. are considered in this study. In addition, the study in [339] proposed a new acoustic feature called the Multi-Resolution Cochleogram (MRCG), which is encoder multi-resolution power distribution in

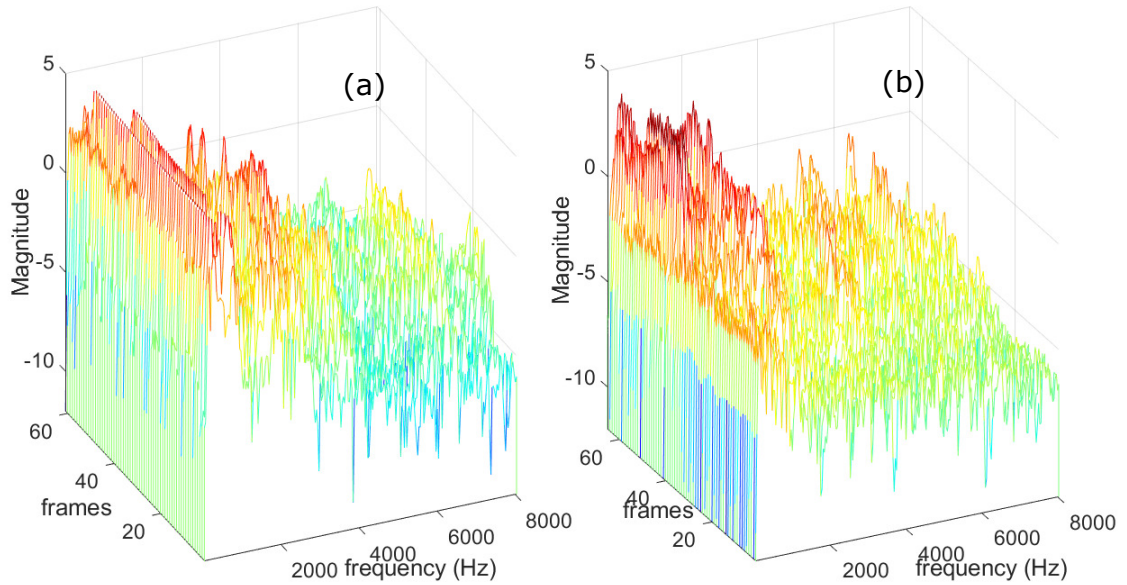


Figure 6.12: Waterfall Characteristics of: (a) Normal Speech, and (b) Dysarthric Speech. After [24].

the time-frequency representation of a signal. Finally, study in [339] found MRCC and pitch (F_0) as complementary features using group Lasso (the least absolute shrinkage selection operator) that improve l_1/l_2 mixed norm regularization on logistic regression to investigate the complementary features. The study in [339] is extended in [340] for monaural speech separation from a supervised learning perspective by predicting an ideal time-frequency mask from the similar acoustic features of noisy speech under reverberant conditions at low SNRs and employing a simple DNNs as a learning machine. The key findings of this study is that complementary feature sets for speech separation in reverberant conditions are different from those in anechoic conditions (as reported in [339]).

Motivated from these studies, we employ such representation for the dysarthric severity-level classification problem. Figures 6.11-(a) and (b) show the plot of STFT *vs.* LP spectrum for the normal *vs.* dysarthria speech case. We also show the waterfall plot in Figure 6.12 to emphasize the corresponding joint time-frequency characteristics during the production of dysarthric speech. From the waterfall plots, we can observe that the formant structure is severely damaged for dysarthric speech as compared to its normal counterpart, where formant peaks and their evolving structures are clearly visible. Thus, the analysis presented in this Section indicates that F_0 , its harmonics, formants, and their structures are severely affected due to dysarthria, more so for high severity and hence, we propose to exploit this unstructured spectral energy distribution captured via spectrograms as feature representation for the proposed deep learning architecture.



Figure 6.13: Schematic Representation of F_0 Detection. Adapted from [38].

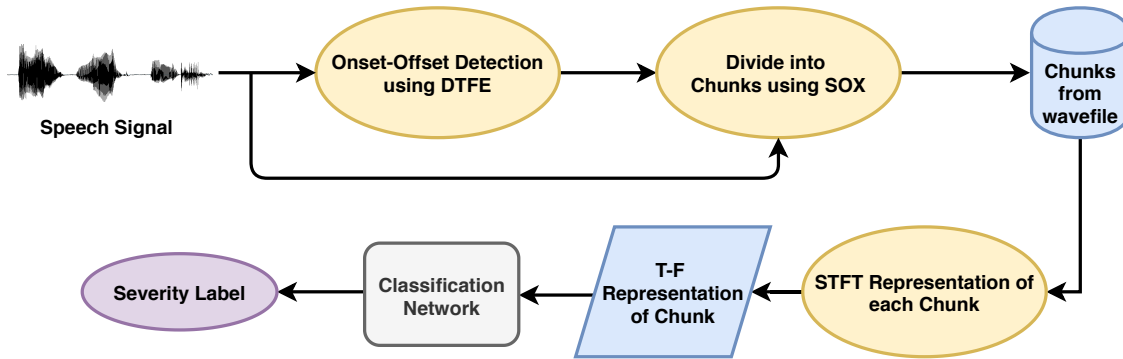


Figure 6.14: Schematic Representation of Proposed Methodology. After [39].

6.4.4 Proposed Approach

In this Section, a detailed description of the methodology and strategies used to solve the proposed problem is provided. Specifically, as represented in Figure 6.14, the following three major components exist:

1. Onset-offset detection;
2. Time-Frequency (T-F) representation of selected short-duration speech segments;
3. Mapping technique for utilizing features to do efficient classification.

6.4.4.1 Onset-Offset Detection

The onset and offset regions of the speech signals were characterized, as a function of their F_0 , by using the Direct Time Fundamental Frequency Estimation (DTFE) method, described in a study reported in [38]. DTFE is a novel algorithm for F_0 estimation performed directly in the time-domain. In this algorithm, F_0 detection is performed via evaluating the actual F_0 candidate from the distance between neighboring significant peaks (i.e., local extrema) that there is only one peak representing the absolute maximum and one the absolute minimum in the quasi-period of the signal. The structure of the F_0 detection is shown in Figure 6.13. Implementation details for pitch (F_0) tracker DTFE are presented here².

²<https://personal.utdallas.edu/~hynek/tools.html> {Last Accessed: July 07, 2022 }

We carefully used that method to extract the onset and offset time stamps from each input speech signal in our dataset. After this, the borders were detected and, in addition, 100 ms from each signal was taken to the left and 100 ms to the right of each border, forming the 200 ms-long signals as “chunks”. Each one of those chunks was modeled by using the STFT, as described ahead.

6.4.4.2 Spectrogram: T-F Representation

STFT was applied to each generated chunk, for T-F representation. To feed the classifier, 2 ms-long frames, shifted 0.5 ms over time, were considered in order to generate a spectrogram image with dimensions of 570×450 pixels. Figure 6.15 shows example spectrograms for different severity-levels of dysarthria. The spectrograms were plotted only for one second-long, i.e., for short-duration speech signals. Observably, the energy distribution across the frames, for speakers with different severity-levels of dysarthria, is significantly unlike. Hence, we hypothesize that those short-duration speech segments are sufficient for the intended classification. To support our hypothesis, we show the experimental results in Section 6.4.5.

6.4.4.3 Mapping Technique: CNN *vs.* ResNet

A recent trend indicates that a high number of stacked layers in neural networks provides better results for classification task in general [341]. Nevertheless, accuracy degrades rapidly after the increment in the number of layers. The reason behind this is the ample training error, instead of overfitting [18]. Moreover, current studies show that deep neural networks are more challenging to train due to overfitting, vanishing gradient, and besides additional issues, as explained in [342–344].

Making CNN models deeper for our task is not an appropriate solution. To overcome the limitation of CNN-based architectures, residual learning-based classifiers, ResNet, were used in [18, 345]. The former technique is used as a baseline for comparisons. Although ResNet uses convolution layers as its building blocks, it is more effective than CNNs for image classification, as suggested in [18]. Thus, we decided to use the strength of ResNet for our classification problem, analyzing its results. In residual learning, $f(y)$ is the underlying function, as shown in Section 6.4.2, to be learned by a regular neural network-based classifier, where y is the set of input features, i.e., spectrogram image of the chunk in our case. Due to the network non-linearity, it is capable of learning $f(y) - y$ along with $f(y)$, forcing the classifier to optimize the residual function $F(y) = f(y) - y$. Hence,

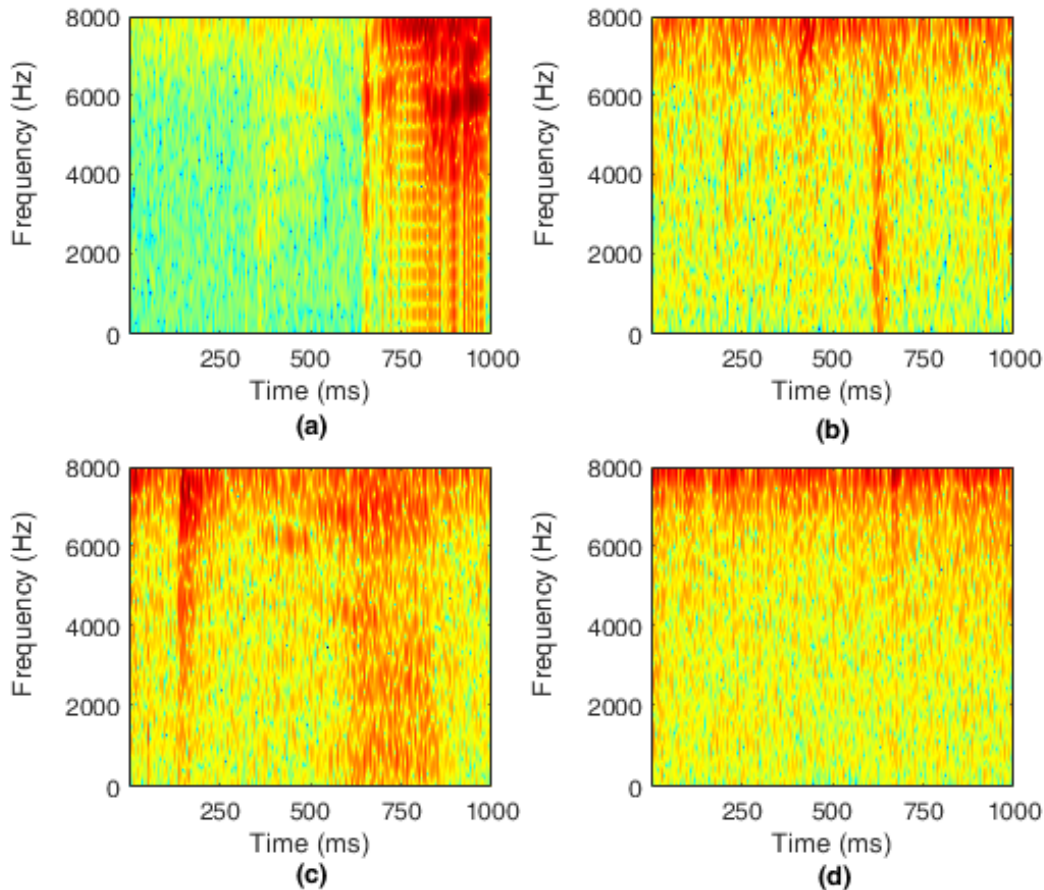


Figure 6.15: STFT of the One Second of Speech Segment of Speakers with Different Severity-Levels, When They Pronounce the Word “Command”: (a) Very Low, (b) Low, (c) Medium, and (d) High. After [24].

the original optimization function becomes $F(y) + y$. Although both the methods are learning the same underlying functions, the ease of learning is different. The main reason behind this is the associated *identity mapping*, as shown in [346]. Due to the skip connections in ResNet, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers. This helps ResNet in learning different patterns more efficiently, as shown in [18,345].

6.4.5 Experimental Setup and Results

In this Section, the description about the database used for experiments is provided, and the details of hyperparameters used both in the baseline system and in the proposed ResNet model. The results for the severity-based classification task for different classifier architectures (GMM, CNN, LCNN, and ResNet) are discussed along with the analysis of the effectiveness of ResNet over the other

Table 6.8: Proposed Architectural Details of ResNet. Here, Conv1 and Conv2 show Continuous Layers of Residual Block, and Conv3 shows Parallel Down-sampling Layer in Residual Block. After [24].

Block	# of Neurons	General Settings	Conv1 Settings	Conv2 Settings	Conv3 Settings
Convolution	3200	64, 7x7, 1	-	-	-
Batch Normalization	-	64,-,-	-	-	-
Max Pool	-	64,7x7,1	-	-	-
Residual Block	1280	-	64,3x3,1	64,3x3,1	-
Residual Block	1280	-	64,3x3,1	64,3x3,1	-
Residual Block	1280	-	64,3x3,1	64,3x3,1	-
Residual Down Sampling	3840	-	128,3x3,2	128,3x3,1	128,3x3,2
Residual Block	2560	-	128,3x3,1	128,3x3,1	-
Residual Block	2560	-	128,3x3,1	128,3x3,1	-
Residual Down Sampling	7680	-	256,3x3,2	256,3x3,1	256,3x3,2
Residual Block	5120	-	256,3x3,1	256,3x3,1	-
Residual Block	5120	-	256,3x3,1	256,3x3,1	-
Residual Down Sampling	15360	-	512,3x3,2	512,3x3,1	512,3x3,2
Residual Block	10240	-	512,3x3,1	512,3x3,1	-
Residual Block	10240	-	512,3x3,1	512,3x3,1	-
Average Pool	-	512,8x8,-	-	-	-
Fully-Connected Layer	4	-	-	-	-

methods for our classification task.

6.4.5.1 Dataset

Universal Access (UA) corpus [23] was used in our experiments. This dataset includes details on speech intelligibility for each dysarthric speaker, in terms of severity-level, based on transcription tasks at the word-level performed by the human listeners. In our experiments, we used 8 speakers, i.e., 4 males, namely, M01, M05, M07, M09, and 4 females, namely, F02, F03, F04, F05. Details about them can be found in [23]. Each speaker produced a total of 765 isolated words, in which 455 words are distinct. For training and testing, we used 90 % and 10 % data, from 455 distinct words for each speaker, respectively.

6.4.5.2 Comparison Methods

Our ResNet model has two types of residual blocks: (i) regular residual block; and (ii) downsampling-based residual block [18]. In this work, our ResNet structure comprises nine regular and three downsampling-based residual blocks. In residual blocks, we used two 2-dimensional CNN layers with a kernel size of $3 \times 3 = 9$. However, for downsampling-based residual blocks, we increase the stride to 2 for the first CNN block and, before adding input x to the output of second CNN block, we use downsampling with a similar setting as the first block. Therefore, we use one downsampling-based shortcut connection with two residual blocks to process the downsampled output. Table 6.8 shows the architectural details of the

proposed model. We first used a single CNN layer with 7×7 kernel size to down-sample the input. Later, we used a total of 14 different residual blocks and, at the end, we adopted a single fully-connected layer with softmax activation function to predict the severity of the input dysarthric speech spectrogram. Architectural details related to the proposed ResNet are shown in Table 6.8.

For the baseline system, we have used GMM as a classifier [174, 212, 347, 348] (Chapter 3). The CNN-based architecture consists of a 5×5 kernel for each one of its four CNN layers: CNN-layer-A, CNN-layer-B, CNN-layer-C, and CNN-layer-D, with 8, 16, 32, and 64 output channels, respectively [349]. Moreover, we adopted max-pooling with a kernel size of 4×4 after the first three CNN blocks. Later, we used three fully-connected layers with 128, 64, and 4 output neurons. ReLU was used as an activation function for the hidden layers in both the models. Accordingly, the output layers in both the models are followed by a softmax activation function. The models were trained for 30 epochs with learning rate of 0.0001, by using Adam optimizer [206]. The LCNN architecture was also employed as studied in Chapter 3. The details of LCNN architecture for this task is shown in Table 6.9.

As described in Section 6.4.4.1, 200 ms-long chunks were extracted from each speech signal. For our experiments, we selected a different number of onset-offset detection routines and then used them for training. In case the distance between consecutive onset-offset tags is less than 200 ms, we used overlapped chunks. In the case of non-overlapping chunks, we got $[(\text{number of chunks}) \times (200 \text{ ms})]$ seconds of speech segment, which becomes, however, less than this in the case of overlapping scenarios. Hence, we took a maximum of $[(\text{number of chunks}) \times (200 \text{ ms})]$ seconds of speech from each utterance for training. The proposed system was assessed for different number of chunks, i.e., different speech duration, in order to prove our hypothesis that maximum one second of speech (i.e., five chunks) is a sufficient time for an efficient classification.

6.4.5.3 Performance Evaluation

Accuracy and F1-score were used for performance evaluation, where the former is the number of correctly predicted wave files out of all the input wave files, and the latter is calculated by taking the harmonic mean of precision and recall for each class. Precision was considered as being the fraction of correct classified instances among all the classifications for each class, and, in addition, recall was defined as being the fraction of correct classified instances among the ones that actually belong to that class. In particular, we calculated the F1-score for each class

Table 6.9: Details of the Proposed LCNN Architecture for Dysarthria Severity Classes. After [24].

Layer	Filter/Stride	Output	#Parameters
Conv1	5x5/1x1	32 x 450 x 570	2432
MFM1	-	16 x 450 x 570	-
MaxPool1	2x2/1x2	16 x 225 x 285	-
Conv2a	1x1/1x1	32 x 225 x 285	544
MFM2a	-	16 x 225 x 285	-
Conv2b	3x3/1x1	64 x 225 x 285	9280
MFM2b	-	32 x 225 x 285	-
MaxPool2	2x2/1x2	32 x 112 x 142	-
Conv3a	1x1/1x1	64 x 112 x 142	2112
MFM3a	-	32 x 112 x 142	-
Conv3b	3x3/1x1	128 x 112 x 142	36992
MFM3b	-	64 x 112 x 142	-
MaxPool3	2x2/2x2	64 x 28 x 35	-
Conv4a	1x1/1x1	128 x 28 x 35	8320
MFM4a	-	64 x 28 x 35	-
Conv4b	3x3/1x1	64 x 28 x 35	36928
MFM4b	-	32 x 28 x 35	-
MaxPool4	2x2/2x2	32 x 14 x 17	-
Conv5a	1x1/1x1	64 x 14 x 17	2112
MFM5a	-	32 x 14 x 17	-
Conv5b	3x3/1x1	32 x 14 x 17	9248
MFM5b	-	16 x 14 x 17	-
MaxPool5	2x2/2x2	16 x 7 x 9	-
FC6	-	1 x 128	24704
MFM6	-	1 x 64	-
FC7	-	1 x 4	260

and presented the “macro average” results. We analyzed the performance of both systems in terms of different speech duration, i.e., maximum [(number of chunks) × (200 ms)] seconds. From Figure 6.16-(a) and Figure 6.16-(b), we can clearly see that the proposed ResNet-based approach outperforms the baseline CNN. In particular, we got, on average, 21.35 % and 22.48 % of improvement compared with the baseline CNN in terms of classification accuracy and F1-score, respectively. For further comparisons, GMM and LCNN classifiers were also considered.

It can be observed that GMM performed relatively poor compared with the other systems, indicating its unsuitability for classifying severity-level of dysarthria

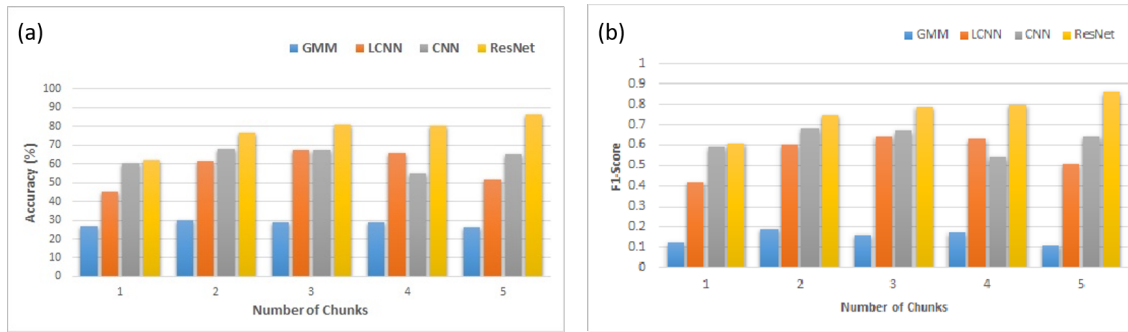


Figure 6.16: Baseline CNN *vs.* ResNet, for Different Speech Duration Based on (a) Classification Accuracy Score, and (b) F1-Score. Additionally, LNCC and GMM were also Considered for Comparisons, however, Since GMM Exhibit a Poor Accuracy, Its F1-Scores were Not Even Computed. After [24].

from short-duration speech. Moreover, GMM is based on the first two moments only, i.e., mean and variance, which may not be adequate to represent nonlinearities in speech production mechanism, and more so, for dysarthric speech, as discussed in Section 3. In addition, estimating higher-order moments with the same statistical confidence, as that of first two moments, requires a large amount of training data, which is not feasible in this problem due to the serious difficulty of getting long-duration dysarthric speech data. Contrary to this, deep learning architectures, in particular the proposed ResNet, is able to capture such nonlinearities from short-duration speech segments.

6.4.5.4 Analysis of Results

In this sub-Section, we analyze how effective the proposed methodology is in two different aspects: (i) learning performance, and (ii) amount of training data. Since CNN and ResNet performed relatively better than the LCNN and GMM, as discussed in the previous sub-Section, we assess hereafter just the behavior of CNN and ResNet-based systems w.r.t. learning performance and the amount of training data.

To analyze the learning performance, we observed the output of the last layer just before the softmax activation from both of our architectures, i.e., CNN and ResNet. To analyze efficiently, we converted the image into binary format, where the white colour part shows the pattern learned by the architecture, as illustrated in Figure 6.17. For that analysis, we used *Guided Backpropagation Saliency* method in order to extract the region learned by any trained CNN-based classifier [350]. In guided backpropagation, forward pass was performed till the target layer on input features is performed. Then, the disadvantageous neurons were kept to

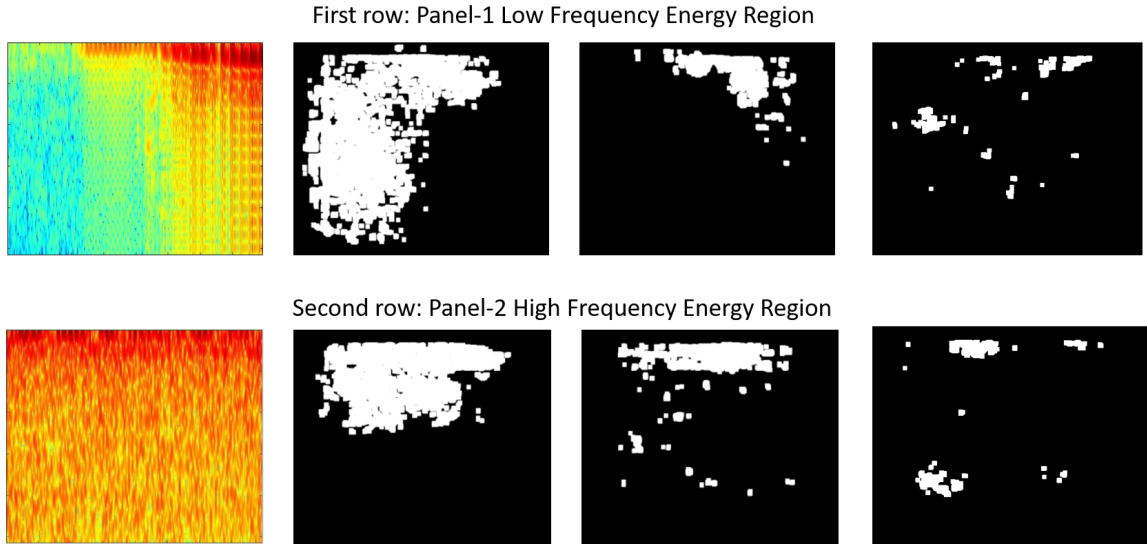


Figure 6.17: Learning of Proposed ResNet *vs.* Baseline CNN. For Both Panels, We Have: [First Column]: Input Spectrogram of Chunk (Horizontal-Axis: Time, Vertical-Axis: Frequency); [Second Column]: Visualization of Learning of ResNet; [Third Column]: Visualization of Learning of CNN; [Forth Column]: Visualization of Learning of LCNN. Here, Visualization Images Are in the Form of Pixels. After [24].

zero and back propagation was applied till the input features. More formally, the entire process can be explained as: [351]:

$$\begin{aligned}
 \text{activation:} & \quad f_i^{l+1} = \text{relu}(f_i^l) = \max(f_i^l, 0) \\
 \text{backpropagation:} & \quad R_i^l = (f_i^l > 0) \cdot R_i^{l+1}, \text{ where } R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}} \\
 \text{backward 'deconvnet':} & \quad R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1} \\
 \text{guided backpropagation:} & \quad R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1} \quad .
 \end{aligned}$$

From Figure 6.17, it can be observed that the advantage of ResNet over CNN for the dysarthric severity-level-based classification, and we can see that ResNet can learn various characteristics of dysarthric speech, which are different from natural speech signal. To understand the advantage of ResNet in our problem, we explored the energy parameter. The energy in dysarthric speech is more distributed, i.e., energy fluctuations are more frequent, compared with the natural speech signal [352,353]. Moreover, it is observed that as the severity-level changes, the energy of dysarthric speech shows significant changes [354]. Hence, capturing these energy fluctuations from the spectrogram is an essential task. In the other words, detecting patterns from low and high frequency regions from the spectrogram can achieve this task and help the model to distinguish between different

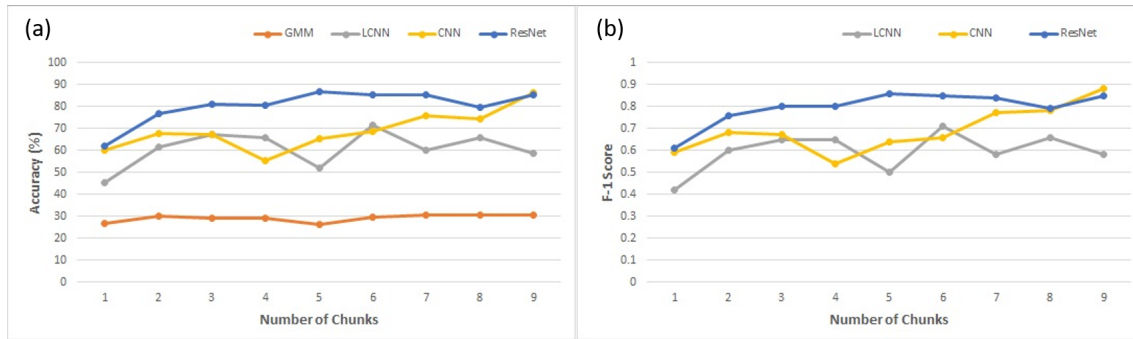


Figure 6.18: Evaluation of Baseline CNN *vs.* ResNet for Different Number of Chunks (i.e., Amount of Training Data) Based on (a) Classification Accuracy Score, and (b) F1-Score. As Previously Presented, GMM and LCNN were Considered for Comparisons, However, Since GMM Exhibit a Poor Accuracy, Its F1-scores Were Not Even Computed. After [24].

Table 6.10: Evaluation of Baseline CNN *vs.* ResNet When Entire Speech Utterance is Available for Training. After [24].

Systems	Accuracy (%)	F1-Score
ResNet	98.90	0.98
CNN	91.76	0.91

severity-levels. Consequently, our short-duration speech segments include both the patterns (i.e., high and low energy regions), as shown in Figure 6.15. From Panels I and II, it can be observed that ResNet is capturing both the regions efficiently and hence, performing better compared to the CNN. In Figure 6.17, CNN captures *only* high energy regions, however, fails to capture low energy regions and hence, performing poor compared to the ResNet.

Here, we analyze the performance of both the systems w.r.t. the amount of training data. To do so, we increased the number of chunks one-by-one, i.e., increasing a speech duration by a maximum of 200 ms, as shown in Figure 6.18. Observably, for five chunks, ResNet performance is high in terms of classification accuracy and F1-score. This analysis empirically supports our hypothesis that one second-long speech segments are sufficient for an efficient classification task. With a maximum of one second-long speech segment, we got 86.63 % classification accuracy and 0.86 F1-score for ResNet. Contrary to this, we obtained 64.35 % classification accuracy and 0.64 F1-score for CNN, respectively.

In complement, as the goal of this work is to detect dysarthria severity-level by using short-segments of speech, we also analyzed both ResNet and CNN structures, when the entire speech utterance is available for training and testing. As shown in Table 6.10, ResNet exhibits superior performance for this task. Hence,

ResNet outperforms baseline CNN for both the classification scenarios, i.e., using short-duration speech segments and using the entire speech utterances. This definitively proves ResNet is superior for the intended classification task.

6.4.5.5 Complementary Comments

Additionally, it can be observed that LCNN could capture both high and low frequency regions, however, the capture ratio is significantly lower compared to the ResNet and CNN, as also shown in Figure 6.17. In contrast, ResNet is efficiently capturing both the regions. Therefore, our ResNet structure not only outperforms the baseline CNN, but also LCNN and GMM. Observably, GMM was the worst classifier in terms of accuracy for this problem.

In addition to the mentioned comparison methods, we have also investigated the other variants of the ResNet architecture, in particular, ResNeSt, which uses a *Split Attention* layer between the two convolutional layers in the ResNet block. However, results obtained were poorer than the original ResNeSt architecture. Therefore, it has not been included in this chapter.

6.5 Classification of Normal *vs.* Pathological Infant Cry

Infant cry analysis and classification is highly interdisciplinary in nature involving physiological, neurological, pediatrics, engineering, developmental linguist, and psychology [355]. Around three million infants die within the first month after the birth, which may be due to vaccine preventable diseases, other pathologies, and malnutrition. In this context, recently fingerprint-based biometrics are developed for infants [356], in addition to cry-based recognition of infants [357]. With respect to various diseases, birth asphyxia and other breathing-related conditions, such as Sudden Infant Death Syndrome (SIDS) are the leading cause of death for newborns [358]. Clinical diagnosis of asphyxia requires analysis of an arterial blood sample of newborns to measure blood gasses, pH, oxygen saturation, and electrolytes, which requires a blood gas - a routine procedure in developed countries, however, in many developing countries it is not, as this procedure is costly and logistics heavy. Hence, asphyxia is generally detected only from emergency and visual symptoms, such as pale and bluish limbs, however, by then severe neurological damage would have already been occurred to the newborns [358, 359]. Similarly, acoustic characteristics of deaf infants depends on hearing loss, type

and period of rehabilitation, and the age of pathology identification [360]. Thus, there is a need to develop a cost effective and non-invasive cry diagnostic tool to assist pediatrics to detect early warning signs of such pathologies. To that effect, this study proposes signal processing-based approaches for infant cry classification task, where asphyxia and deaf cry samples are considered as pathological samples.

The earlier investigations on infant cry analysis can be studied from [355, 361, 362]. The state-of-the-art MFCC feature set along with GMM as a classifier was employed in [363, 364]. In this study, we propose two efficient feature representations for infant cry analysis, namely, CQCC [26] and subband Teager Energy representations [27]. These approaches are explained in subsequent sub-sections.

6.5.1 Form-Invariance Property of CQT

The details of the CQT can be studied from earlier Chapter 5, section 5.2.3.1. For the sake of simplicity, we consider continuous-time version of FT, STFT, and CQT. If $x(t)$ and $X(\omega)$ are Fourier transform-pair, where t and ω represents the time and frequency index, respectively, then time-scaling property of CTFT implies [42, 365]:

$$\mathcal{F}\{x(\alpha t)\} = \frac{1}{|\alpha|} X\left(\frac{\omega}{\alpha}\right), \quad (6.29)$$

and thus, a linear time-scaling corresponds to a frequency scaling by an *inverse* factor of $\frac{1}{\alpha}$ and vice-versa, indicating the *form* of spectrogram is unaffected and hence, the name *form invariance*. However, this property does not hold for the traditional STFT, where analysis window function is dependent *only* on time parameter. In particular, Schroeder and Atal defined the STFT through a practically realizable bandpass filters [250]. In particular,

$$F(t, \omega) = \int_{-\infty}^t f(\tau) w(t - \tau) e^{-j\omega\tau} d\tau, \quad (6.30)$$

where $w(t, \tau)$ represents the analysis window. For form-invariance of STFT, we must have

$$F(t, \omega) = \gamma F(\alpha t, \beta \omega), \quad (6.31)$$

where α and β are scaling factor for time and frequency, respectively. However, it is shown in the literature that realization of eq. (6.31) yields the necessary and sufficient condition on weighting (i.e., window) function, which belongs to the class of single-term power functions, i.e., $w(t) = a \cdot t^b, t > 0$, and as per stability

condition for LTI filter, this filter is unstable and hence, practically not realizable [366]. However, it is interesting to note that if the window function is made to be frequency-dependent, i.e., $w(t) \equiv w(t, \omega)$ (as in the case of CQT). In particular, eq. (6.30) becomes

$$F(t, \omega) = \int_{-\infty}^t f(\tau)w(t - \tau, \omega)e^{-j\omega\tau} d\tau, \quad (6.32)$$

the form-invariance property, i.e, eq. (6.32) is satisfied by eq. (6.33) for the window function, i.e.,

$$w(t, \omega) = v(t, \omega)t^b, \quad t > 0, \omega > 0, \quad (6.33)$$

where $v(t, \omega)$ is an arbitrary real function of (t, ω) , and b is real constant and function $w(t, \omega)$ also satisfy Bounded Input Bounded Output (BIBO) stability condition for LTI filter, i.e.,

$$\int_{-\infty}^{\infty} |w(t, \omega)| dt < \infty. \quad (6.34)$$

Furthermore, eq. (6.33) also holds for window function considered in most practical model and short-time analysis performed by the peripheral auditory system. For example, the original model developed by Flanagan [367] represents the window function for the mechanical spectral analysis due to the movements of BM in the cochlea of human ear [365]. In particular, $w(t, \omega) = (t\omega)^2 e^{-\frac{t\omega}{2}}$, which is similar to eq. (6.33).

6.5.2 Subband Teager Energy Representations

For detection of the pathology using infant cry, TEO-based features are also explored. The details of the TEO can be studied from Chapter 4, Section 4.3. As discussed earlier, TEO is originally derived to estimate the energy for a monotone signal. However, speech signal consists of the frequency range varying from low frequency band to Nyquist frequencies. Hence, in order to obtain the monotone approximation of the signal, speech signal is allowed to pass through the filterbank, which consists of several subband filters with appropriate center frequency and bandwidth. The subband filtered signals are narrowband signals, which are supposed to approximate the monotone signals and hence, TEO can be applied on these subband filtered signals. In this work, Gabor filterbank with linearly-spaced subband filters, is utilized for subband filtering. TEO is applied on each subband filtered signal to accurately estimate the signal's energy. Further-

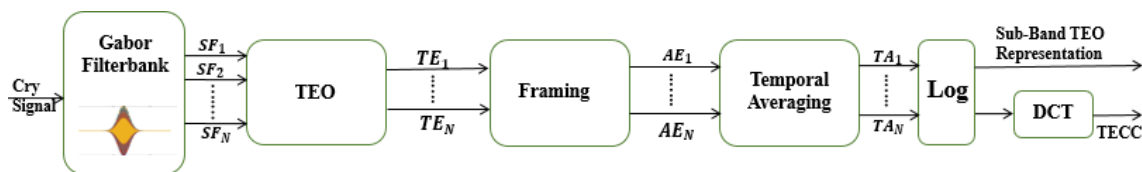


Figure 6.19: Functional Block Diagram of the Proposed Subband TEO Representation and TECC Feature Set. (SF: Subband Filtered Signal, TE: Teager Energies, AE: Averaged Energies over Frames). After [40,41].

Table 6.11: Statistics of the Baby Chillanto dataset. After [25].

Class	Category	# Samples
Healthy	Normal	507
	Hunger	350
	Pain	192
Pathology	Asphyxia	340
	Deaf	879

more, these narrowband energies are segmented into the frames of 20 *ms* duration with overlapping of 10 *ms*. Then, temporal average for each frame is estimated to produce N -dimensional (D) *subband Teager energy representations (subband-TE)*. Discrete Cosine Transform (DCT) is performed on *subband Teager energy representations* to obtain the TECC. The functional block diagram representation of the proposed subband-TE and TECC feature set is shown in Figure 6.19. In this study, we analyzed the relative performance of the subband-TE *vs.* TECC feature set.

6.5.3 Experimental Setup

Standard and statistically meaningful, Baby Chillanto database is used in this work. It was developed by the recordings conducted by medical doctors, which is a property of NIAOE-CONACYT, Mexico [25]. Each cry signal was segmented into one second duration (which represent one sample) and are grouped into five categories. Two groups were formed for binary classification of healthy *vs.* pathology. Healthy cry signals include three categories, namely, normal, hungry, and pain resulting in 1049 cry samples. Pathology cry signals include two categories, namely, asphyxia and deaf resulting in 1219 cry samples. Table 6.11 shows the statistics of Baby Chillanto database. Experiments are performed using 10-fold cross-validation.

Proposed CQCC feature set is employed with 90-dimensions (90- D), which includes static, Δ , and $\Delta\Delta$ features. For fair comparison, we also extracted the 90- D feature sets from STFT, named as *cepstrals*. Furthermore, we used the state-

of-the-art MFCC feature set, extracted from the magnitude spectrum along with Mel filterbank that uses Mel-scaled bandpass filters [368]. In LFCC, Mel-scaled bandpass filters are replaced by linearly-spaced bandpass filters. For, LFCC and MFCC, we preserved initial 13-dimensions (13- D), and then Δ and $\Delta\Delta$ coefficients are appended to it, which makes 39- D feature sets. The proposed subband-TE and TECC feature sets are also extracted using 40 number of subband filters in the filterbank. Furthermore, subband-TE being a spectral representation, its performance is compared against the Mel Filterbank coefficients (MelFB), Linear Filterbank Coefficients (LinFB), and STFT. MelFB and LinFB are extracted using 40 number of subband filters.

For infant cry classification task, we use two state-of-art classifiers, GMM and SVM, which are commonly used for infant cry classification task [355, 363]. The SVM was utilized in [8] and hence, we employed it as a classifier for another baseline architecture. The details of GMM and SVM can be studied in [174]. Furthermore, performance of various systems is evaluated using two performance metrics, namely, % classification accuracy and % EER [179].

6.5.4 Experimental Results using CQCC

6.5.4.1 Spectrographic Analysis

Panel-I and Panel-II shows the waterfall plots and corresponding top view of STFT and CQT for healthy *vs.* pathology cry signal, respectively. Figure 6.20(a) and Figure 6.20(b) shows the waterfall plot of STFT and its top view, where it can be observed that F_0 of the normal signal occurs above 300 Hz. Whereas, for pathological cry signal, the anomaly in the cry signal, which appears like F_0 , is estimated in the lower frequency regions. From Figure 6.20(c) and Figure 6.20(d), it can be observed that CQT emphasizes this anomaly in a much better way due to its high frequency resolution for lower frequencies.

6.5.4.2 Results using Evaluation Metrics

Table 6.12: Results in % Classification Accuracy (Acc) for Various f_{min} (Hz) of using GMM. After [26].

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

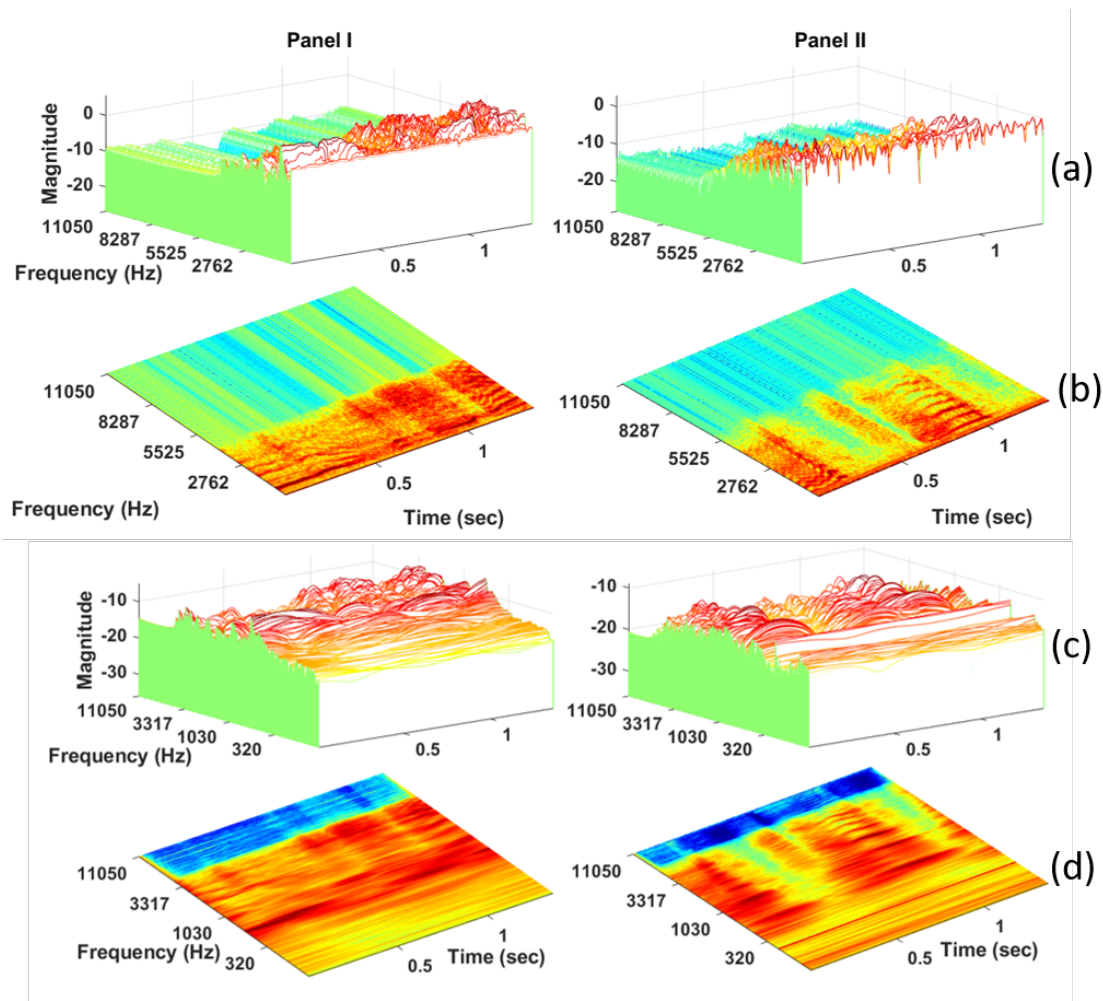


Figure 6.20: Panel-I and Panel-II Depicts the Spectrographic Analysis for Healthy (Normal) and Pathology (Asphyxia) Infant Cry Signal: (a) the Waterfall Plot for STFT, (b) the Top View of the STFT Waterfall Plot, (c) Waterfall Plot for CQT, and (d) the Top View of the CQT Waterfall Plot. After [24].

All the experiments in this work are performed using 10-fold cross-validation. Initially, we performed the experiments by varying the f_{min} in eq. (5.22) with hann window in CQT and keeping the 512 number of Gaussian mixtures. It can be observed from Table 6.12 that best possible results are obtained with $f_{min} = 100$ Hz. This might be due to the fact that infant cry generally consists of F_0 above 250 Hz and hence, $f_{min} = 100$ Hz would be the optimum choice to capture the anomaly in the infant cry signal. Furthermore, experiments are performed w.r.t. various analysis window by keeping the 512 number of Gaussian mixtures in GMM as shown in Table 6.13. It can be observed that relatively better results are obtained using hann window. Furthermore, we also analyzed the performance w.r.t. number of Gaussian mixtures in GMM, and it is observed from Table 6.14 that 512 Gaussian mixtures are more suitable to estimate distribution of this data.

Table 6.13: Results (in % Classification Accuracy) for Various Window Functions using GMM. After [26].

Window	Acc.	Window	Acc.
Hann	99.82	Hamming	99.60
Gaussian	98.81	Rectangular	97.75

The experimental results obtained (in % classification accuracy and % EER) using combination of the various feature sets along with the GMM and SVM classifiers are reported in Table 6.15. It can be observed that relatively better performance is obtained for the proposed CQCC feature set using both GMM and SVM classifiers. Furthermore, it can be observed that CQCC and MFCC performs better than the *cepstrals* and LFCC, respectively. Here, MFCC and CQCC feature sets are designed w.r.t. perception of sounds in human auditory systems, which uses non-linear (in particular, logarithmic) scale along frequency-axis. Hence, we can conclude that the human auditory system-based features performing better as compared to the linear-scale features for the pathology cry detection. The sim-

Table 6.14: Results (in % Classification Accuracy) *w.r.t.* Number of Mixtures. After [26].

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

ilar trends in results, as that of in Table 6.15, are observed in DET curves (having discontinuities due to less number of trials because of insufficient data) shown in Figure 6.21. Furthermore, the performance of the proposed feature set is also validated by performing the standard statistical testing. To that effect, we have performed the 10-fold cross-validation experiment for 50 times for each feature set, and it was observed that the mean and median values of % classification accuracy for CQCC feature set are better than MFCC and LFCC feature sets, indicating statistical significance of proposed CQCC feature set. On the whole, proposed CQCC feature set performs better than the existing features for various evaluation factors, may be due to presentation of *form-invariance* property and CQT so that CQCC as feature descriptors is able to represent discriminative features of normal *vs.* pathological infant cry.

Table 6.15: Results in (% Classification Accuracy and % EER) for Various Feature Sets using GMM as a Classifier. After [26].

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

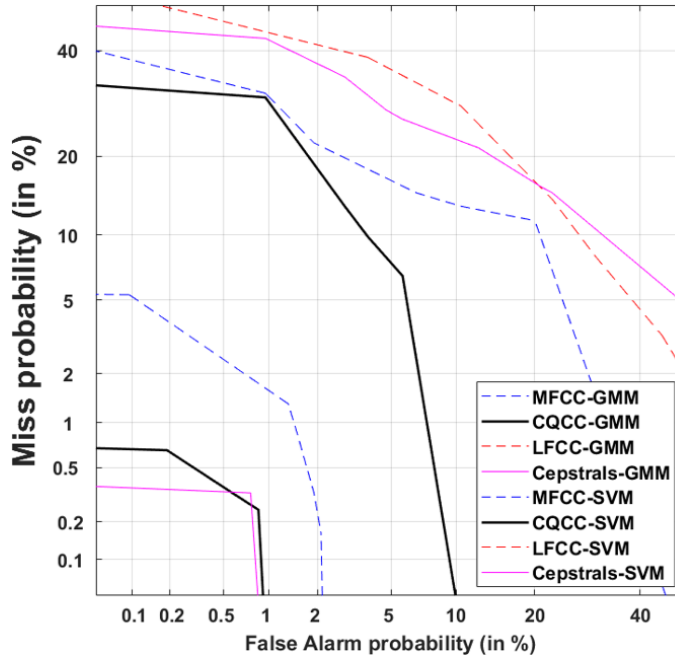


Figure 6.21: DET Curves Obtained for Various Features using GMM and SVM Classifiers. After [26].

6.5.5 Experimental Results using Subband-TE Features

6.5.5.1 Spectrographic Analysis

In Figure 6.22, Panel-I and Panel-II represents the spectrographic analysis for randomly sampled normal and asphyxia cry signals, respectively. Figure 6.22(a), Figure 6.22(b), and Figure 6.22(c) represents the STFT, MelFB, and subband-TE representations, respectively. It can be observed from Figure 6.22(a) that there is a difference in the pattern formed by F_0 and its harmonics for normal *vs.* asphyxia cry signals. These differences in the pattern are also visible for MelFB representation as shown in Figure 6.22(b). However, these differences are more vivid for subband-TE representations as shown in Figure 6.22(c). It might be because of the fact that TEO can accurately estimate the energy of the signal considering non-linear aspects of the speech production mechanism and also properties of airflow

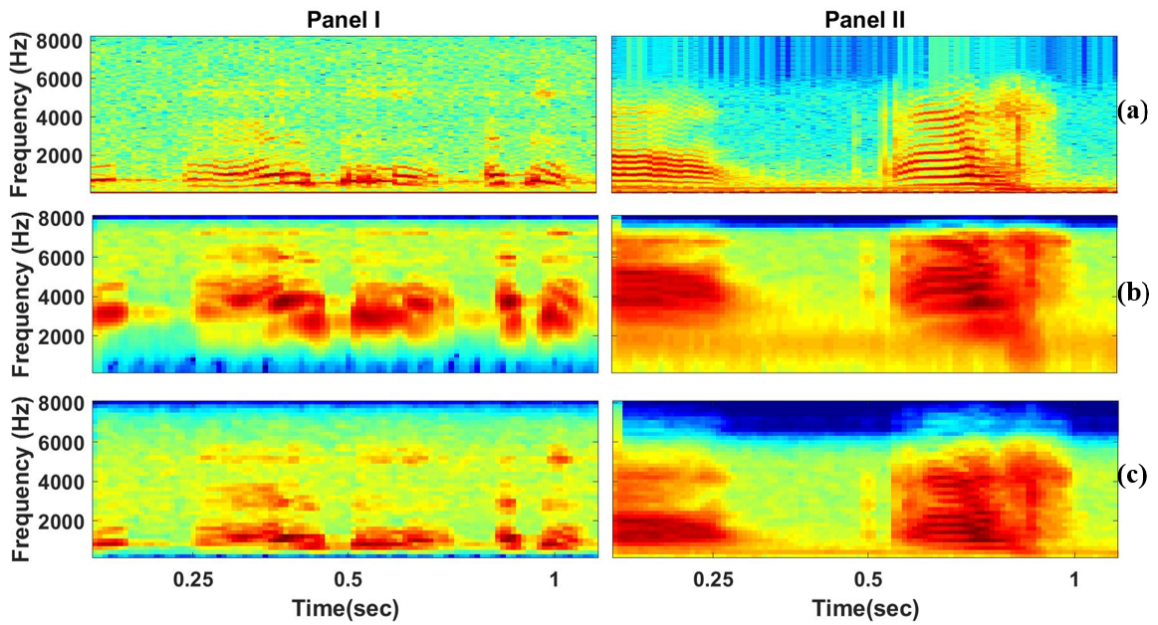


Figure 6.22: Panel-I and Panel-II Represents the Spectrographic Analysis for Normal *vs.* Asphyxia Cry Samples, Respectively. Figure 6.22(a), Figure 6.22(b), and Figure 6.22(c) Represents the STFT, MelFB, and Subband-TE Representations, Respectively. After [27].

pattern in the vocal tract system [42,366]. Furthermore, the results obtained using 10-fold cross-validation also validates that the proposed TECC and subband-TE representations performs better over the other feature sets in this study.

Table 6.16: Results in (% Classification Accuracy and % EER) using Various Cepstral Feature Sets using GMM and SVM as Classifiers. After [27].

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

Table 6.17: Results in (% Classification Accuracy and % EER) for Various Spectral Feature Sets using GMM and SVM as Classifiers. After [27].

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

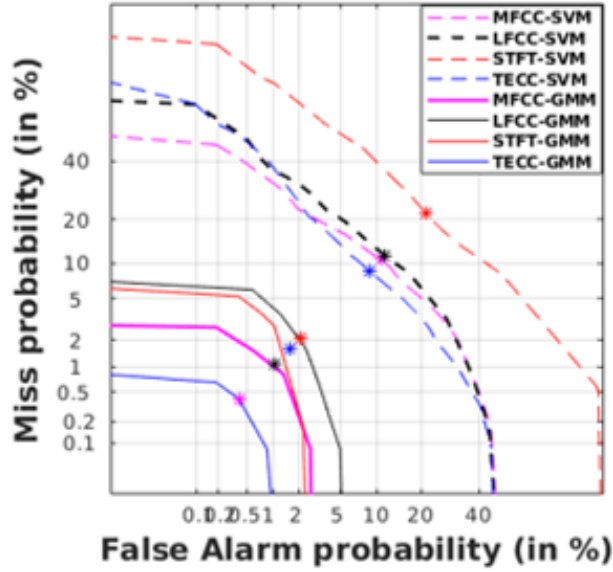


Figure 6.23: DET Plots for Various Feature Sets using GMM and SVM as Classifiers. After [27].

6.5.5.2 Results using Evaluation Metrics

Cepstral representations are being common in speech signal processing applications, we performed the experiments using four cepstral feature sets, namely, MFCC, LFCC, STCC, and TECC. As the size of the dataset is relatively small, experiments are performed using 10-fold cross-validation. The dataset consists of the healthy and pathology class cry samples recorded with sampling rate of 22 kHz and 11 kHz, respectively. The experiments are performed using features extracted from the cry samples resampled to 16 kHz and results are reported in Table 6.16. It can be observed that the proposed TECC feature set outperforms the other feature sets for both SVM and GMM classifiers. We utilized 512 Gaussian mixtures in the GMMs. Furthermore, experiments are extended with spectral feature sets, namely, subband-TE, MelFB, LinFB, and STFT. We utilized the spectral feature representations as it has low-dimensional representations than the cepstral features. It can be observed from Table 6.17 that the proposed subband-TE feature set outperforms the other feature sets for both SVM and GMM classifiers. Furthermore, all the *spectral* representations performs equally well as compared to their corresponding *cepstral* representations. However, subband-TE performs slightly better than its cepstral counterpart, i.e., TECC. Hence, it would be better to choose the spectral representations for this application.

Furthermore, DET curves are plotted for various spectral features as shown in Figure 6.23. It can be observed that the proposed subband-TE representation performs better than all the other spectral representations for both the classifiers.

The experiments are extended for varying number of Gaussian mixtures in GMM and results are obtained as shown in Table 6.18. It can be observed that the performance is improving as we increase the number of Gaussian mixtures in GMM from 64 to 512 and then it saturates, possibly due to the fact that a large number of 1024 mixtures is not required to model relatively lesser duration of infant cry samples. Hence, we utilized 512 Gaussian mixtures in GMM for the remaining experiments. Furthermore, performance is also validated w.r.t. number of subband filters in the Gabor filterbank to extract the subband-TE representations, and the results are reported in Table 6.19. It can be observed that the performance is almost constant w.r.t. number of subband filters in the filterbank and hence, we chose 40 subband filters in the filterbank as an optimal choice.

Table 6.18: Results (in % Classification Accuracy) *w.r.t.* Number of Mixtures. After [27].

# Mixtures	64	128	256	512	1024
Accuracy	98.72	98.94	99.16	99.47	99.47

Table 6.19: Results in % Classification Accuracy (Acc) for Various Number of Filters using GMM. After [27].

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

Based on the experimental results, we investigated the suitability of the spectral representations over cepstrals for infant cry analysis and classification. Because of the high pitch-source harmonics, spectral representations are more suitable for the normal *vs.* pathological infant cry classification. This theoretical assumption is validated using the experimental results. Furthermore, we exploited the capability of the TEO for accurately estimating the energies (especially approximated for the lower frequency regions). TEO being capable of better approximating the energies in low frequency regions, it is the suitable choice to extract information for pitch-source harmonics of infant cry, which is present at low as well as higher frequency regions of the spectrogram.

6.6 Chapter Summary

In this chapter, speech applications other than the feature development for anti-spoofing, are discussed. It includes the investigation on significance of CMVN

for replay SSD task, significance of the DAS *vs.* MVDR beamformer techniques for replay SSD task on VAs, suitability of ResNets for severity-level classification of the dysarthric speech, and feature sets (CQCC and subband-TE) for infant cry classification.

In the study of CMVN technique, we presented analysis on ASVSpooof-2017 and -2019 challenge datasets. The contradictory behavior is observed for the application of CMVN on these two datasets. For ASVSpooof 2019 dataset, it is observed that the *pdfs* of the individual replay configurations are lying on the one of the side than that of bonafide *pdf* for without CMVN case. Furthermore, the *pdfs* of the replay configurations are separated apart from the *pdf* of the bonafide data, as the replay configuration characteristics are intensified. By applying CMVN, these replay characteristics are suppressed and bringing the *pdfs* of bonafide and spooof data closer to each other, which loses the classification capability to a certain extent. This might be because of the generation of the replayed speech samples by simulating the acoustic and replay configurations, rather real replay environments. However, in ASVSpooof 2017 challenge dataset for without CMVN case, *pdfs* of the individual replay environments are lying on both the sides of the *pdf* of the bonafide data. Hence, the cumulative effect of all the replay environments might create difficulty for the classification of bonafide *vs.* spooof speech. However, the results are similar for the application of CMVN in environment-dependent cases for both the datasets. Finally, we observe that the applicability of the CMVN on cepstral features for the classification task depends upon the intended dataset, which can be analyzed using the *pdfs* of the sample data.

In beamforming analysis study, significance of DAS beamformer over MVDR for replay SSD task on VAs is analyzed. This crucial observation found in this work is contradictory w.r.t. suitability of state-of-the-art MVDR beamformer for DSR, indicating straightforward generalization of beamforming method from DSR to replay SSD in VAs is not recommended even though DSR is very much integral part of VAs. In addition, due to linear phase characteristics of DAS beamformer, the acoustical characteristics of reverberation in replay spooof are presented and hence, TECC is employed to capture these reverberation characteristics. Performance comparison with the existing CQCC and LFCC indicates better performance offered by TECC. Our future work will be directed to extend this work on the other beamforming techniques, with the aim of capturing reverberation.

In the third component of this chapter, a novel technique to detect dysarthria severity-levels was proposed. In particular, we presented time-domain and frequency-domain analysis of dysarthric speech to justify spectrogram as feature represen-

tation particularly capable of capturing unstructured spectral energy density distributions. Our results indicate that GMM perform poorer than the other systems, suggesting deep learning-based architectures and, in particular, the proposed ResNet. Based on short-duration speech segments and ResNets, our strategy differs from current state-of-the-art methods, in which long-duration speech segments are feed to a CNN. Our relevant experiments show that the former classifier outperforms the latter in terms of accuracy and F1-score, not only for short-speech segments but also for long ones. We observed, however, that only ResNet succeed in using short speech tags to detect severity-levels of dysarthric speech.

The fourth component of this chapter consists of the analysis of the CQCC and TEO-based features for infant cry classification task. Experiments are performed with various cepstral features, such as MFCC, LFCC, STFT-based cepstrals, CQCC, and TECC feature sets. Furthermore, experiments are also performed with spectral-based features, such as MelFB, LinFB, STFT, and subband-TE. Among these feature sets, CQCC and subband-TE performs relatively better over the other feature sets. We believe that CQCC feature set preserve the *form-invariance* property thereby making feature descriptors invariant w.r.t. linear scaling and hence, preserve discriminative features of normal *vs.* pathological infant cries. TEO being capable of better approximating the energies in low frequency regions, it is also the suitable choice to extract information for pitch-source harmonics of infant cry, which is present at low as well as high frequency regions of the spectrogram.

CHAPTER 7

Summary and Conclusions

This chapter describes the summary of the entire thesis work along with the limitations of the current work, potential future research directions, and a few open research problems.

7.1 Summary of the Thesis

This thesis work is briefly introduced in Chapter 1. The literature survey is presented in Chapter 2. In Chapter 3, various components of experimental setup that are extensively used in different experiments, are discussed.

In this thesis, various CM approaches for ASV and VAs are developed against three major spoofing attacks, namely, replay, SS, and VC. Furthermore, the presence of the pop noise in the live speaker is also considered for replay SSD task. In particular, the development of the handcrafted features for anti-spoofing is the key contribution of this thesis work. To that effect, various subband filtering-based and spectral-based feature sets are developed. The proposed subband filtering-based feature sets exploit the energies derived from TEO-based frameworks. It includes ETECC, CTECC_{max}, and CFCCIF-ESA feature sets, which are studied in Chapter 4. Whereas, spectral-based feature sets includes the development of CQT for VLD and SRCC for replay SSD task. These spectral representation-based feature sets are studied in Chapter 5. Furthermore, in Chapter 6, these feature sets are explored for the other speech technology applications, namely, severity-level classification of dysarthric speech and infant cry analysis and classification. Chapter 6 also includes other related work on anti-spoofing, namely, analysis of CMVN and beamforming approaches for replay SSD task.

ETECC feature set is developed using the concept of ETEO, which uses the concept of signal mass to get a more precise estimate of signal energy as compared to the traditional TEO. In particular, the TEO-related approximation $\sin(\omega) \approx \omega$ holds true only for lower frequencies and hence, it is not suitable for higher

frequency contents of signals. The concept of signal mass in ETEO compensates the energy in the high frequency regions to provide a more precise estimate of signal’s energy. Subband filtering was performed using Gabor filterbank with linearly-spaced frequency responses. Subband filtering helps to approximate the subband filtered signal to a monocomponent signal, which helps for the accurate estimation of the energies. Furthermore, PFE analysis on ASVSpooof 2017 dataset is also performed for the feature sets in this study. The PFE analysis quantifies the inter-class dissimilarity and intra-class similarity for the various feature sets in the study and gives the confidence for implementing the CM systems. The extensive set of experiments are performed for parameter tuning of the proposed ETECC feature set. Furthermore, the experiments are extended to compare the performance of the state-of-the-art feature sets. The relatively better performance is observed on ASVSpooof 2017 and ReMASC datasets over the other feature sets.

In CTECC_{max} feature set, the multi-channel (and spatial diversity) information in microphone array is exploited for the replay SSD in VAs. To that effect, we provide the mathematical analysis for choosing the appropriate subband channel information (in particular, maximum noise distortion including acoustic reverberation due to replay attack) among the multiple subband channels obtained from the microphone array. The appropriate subband channel information is based on *maximum* cross-Teager energy (an opposed to minimum cross-Teager energy as in the speech recognition literature) estimation among the subband channels, to derive the proposed CTECC_{max} feature set. The experiments are performed using ReMASC dataset. In replay SSD, it is necessary to emphasize the acoustic effects and hence, we chose *maximum* cross-Teager energy to extract these acoustic effects. The proposed CTECC_{max} feature set outperforms the results reported in recently proposed complex deep learning-based architecture and other state-of-the-art feature sets commonly used in the anti-spoofing literature. One of the limitations of ReMASC dataset is absence of well known data partition that is universally accepted (for example, we followed data partition *w.r.t* study reported in [12]) and then, there is need to address this in the near future.

The performance of the CFCCIF-ESA feature set is evaluated on ASVSpooof 2015 dataset, which considers the SS- and VC-based spoofing attacks for ASV system. The discriminative acoustic cue for SS- and VC-based attacks lies in the presence of the artifacts in synthesized and voice-converted speech signals, wherein the speech signal is generated using only magnitude information of the spectrum, neglecting the phase component during signal reconstruction. Thus, phase information in those speech signals is not as natural as in genuine speech signals. This

fact is analyzed by visualizing the IFs from genuine and synthetic spoof speech signals. The proposed CFCCIF-ESA feature set combines the implicit information from magnitude envelopes and IFs estimated using ESA, from the subband filtered signals. The cochlear filterbank is utilized in the subband filtering. In this work, IFs are estimated using ESA, which have relatively low computational complexity, high time resolution, and instantaneously adapting nature, as compared to the Hilbert transformed-based approach that has poor time resolution, and requires computationally complex task of phase unwrapping. The capability of the ESA is reflected into better performance for SSD task. Furthermore, it can be observed that, our proposed CFCCIF-ESA feature set shows significantly better performance for S10-attack, which is known to be most difficult attack to detect for the other feature sets reported in the anti-spoofing literature [2].

Furthermore, CQT-based algorithm is employed to detect the liveness in the genuine speaker by using the pop noise as a discriminative acoustic cue. The experiments are performed on recently released POCO dataset during INTER-SPEECH 2020. The results of the proposed approach are compared against the baseline, where feature sets are derived from the traditional STFT. The VLD systems for the proposed CQT-based algorithm *vs.* STFT-based baseline are developed using various classifiers, namely, SVM, GMM, CNN, LCNN, and ResNet. The relatively best performance is obtained by CQT-based algorithm along with LCNN architecture among all the VLD systems considered in this study. Furthermore, the SRCC feature set is employed for replay SSD task. We investigated physics of replay attack and spectral root cepstrum, where logarithmic nonlinearity in state-of-the-art MFCC is replaced by power-law nonlinearity for replay SSD in the context of VAs. For power-law nonlinearity, dynamic behavior of the output does not depend critically on the input amplitude. A proper choice of γ in SRCC feature extraction plays a vital role in deconvolving the input signal. The selected γ value also pointed out that this system possess more zeros than the poles. The experiments are performed on ASVSpooF 2017 and ReMASC datasets using MSRCC and PSRCC feature sets. For ASVSpooF 2017 dataset, MSRCC and PSRCC feature sets extract complementary information and hence, their score-level fusion produce significant improvement in results. However, ReMASC dataset shows the better performance with MSRCC feature set alone.

Chapter 6 begins with analysis of CMVN technique for replay SSD task. To that effect, analysis on ASVSpooF-2017 and -2019 challenge datasets is presented. The contradictory behaviour is observed for the application of CMVN on these two datasets. It suggests that CMVN acts as double-edged sword, which should

be carefully utilized by analyzing the intended task and dataset used. In addition, significance of DAS beamformer over MVDR for replay SSD task on VAs is analyzed. The crucial observation found in this work is contradictory w.r.t. suitability of state-of-the-art MVDR beamformer for DSR, indicating straightforward generalization of beamforming method from DSR to replay SSD in VAs is not recommended even though DSR is very much integral part of VAs. Furthermore, a novel technique to detect severity-levels of dysarthria was proposed. In particular, we presented time-domain and frequency-domain analysis of dysarthric speech to justify spectrogram as feature representation particularly capable of capturing unstructured spectral energy density distributions. ResNet architecture is employed as a classifier, which shows the relative performance improvement over GMM, CNN, and LCNN classifiers. For infant cry classification, CQCC and TEO-based features are analyzed using GMM and SVM classifiers. Experimental results suggest that CQCC and subband-TE representations are more suitable for this task over the MFCC, LFCC, STFT-based cepstrals, MelFB, LinFB, and STFT feature sets.

7.2 Limitations and Future Research Directions

Limitations and future scope of the work presented in this thesis or in the anti-spoofing literature are as follows:

- For ASV, MFCC feature set continues to be the state-of-the-art feature set, whereas it is not the case for CM for SS, VC, and replay. Thus, we cannot use the same feature set that is used in ASV system to detect spoofed speech; indicating a genuine need of a separate SSD system in tandem with ASV system. Thus, it is necessary to create a joint protocol for evaluation of SSD system in tandem with ASV similar to the performance measure of t -DCF [369].
- The replay spoof detection in this thesis or most of the anti-spoofing studies considers the reverberation effect due to replay spoof mechanism as the main acoustic signature (or cue). However, this is not the case if replay attack is built in outdoor environment. The similar issue was observed during the experiments performed on environment-independent scenario for vehicle environment on ReMASC dataset. It was observed that the SSD system trained on indoor environment fails to identify the spoofing attacks built in vehicle (outdoor) environment. This issue is being practically significant and needs to be analyzed via development of the suitable CM in the future.

- The replay spoof detection in VAs will be the major concern as the utility of VAs is exponentially increasing. VAs utilizes the microphone array and relatively lesser work is reported for SSD task in VAs. In this thesis, we exploited the microphone array for replay SSD task by developing the multi-channel-based $CTECC_{max}$ feature set. The analysis on conventional beamforming techniques for replay SSD task is also provided in this thesis. However, there is still a scope for further improvement. Hence, more efficient feature development and beamforming strategies, especially for replay SSD task can be investigated in the future.
- This thesis also proposed the approach of VLD for anti-spoofing, where proposed CQT-based feature set performs better over previously proposed STFT-based feature set. However, there is still a scope for further improvement by utilizing more efficient pop noise detection methods (improved signal processing or probabilistic approaches) and sophisticated deep learning architectures, in particular, by exploiting various loss functions in CNN, LCNN, and ResNet. Furthermore, the proposed approach do not address the issue of artificially added pop noise in the spoof speech signal, which can be easily added by the attacker at the arbitrary locations in the utterance. Thus, the VLD system can be further modified to detect the pop noise at pop noise-specific phonemes and improve the security for the ASV system, which to the best of author's knowledge is an open research problem.
- The proposed SRCC feature set is independently extended for magnitude and phase components to develop MSRCC and PSRCC, respectively. The score-level fusion of the MSRCC and PSRCC could successfully capture the complementary information in magnitude and phase components of the spectrum for ASVspoof 2017 dataset. However, PSRCC failed to capture complementary information on the ReMASC dataset. Hence, the appropriate signal processing method for phase component can be developed to let it capture the complementary information. Furthermore, in SRCC feature set, γ value is chosen empirically. The suitable strategy for automatic selection of γ value in SRCC framework can be developed in the future.
- The recent study in [370] demonstrate the decomposition of the CQT spectrum into an energy-normalized pitch component and a pitch-normalized spectral component, from which a number of harmonic coefficients are extracted. It results in Constant-Q Harmonic Coefficients (CQHC), which provide a compact and interpretable feature for characterizing the timbre of a

musical instrument. The CQCC, derived from CQT, is state-of-the-art feature set to alleviate various spoofing attacks. However, application of the CQHC for the SSD task is nil and hence, it can be the potential future research direction.

- In the literature and correspondingly in this thesis, linear kernel is utilized for the VLD task, and it gives reasonably good performance. The similar observation is noticed for the classification of the normal *vs.* pathological infant cries. Exploring different kernel functions for SSD or infant cry classification would be a potential future research direction.
- Recently, wavelet signal processing-based features are utilized for VLD [371]. It can be extended in future for SSD in ASV and VAs by fine-tuning the wavelet-based feature set.
- In this thesis, the proposed CFCCIF-ESA feature set combines the magnitude and phase information by multiplication of magnitude with IFs. This idea is motivated by the study reported in [232]. However, other fusion techniques, such as concatenation of the magnitude and phase-based representations or decision-level fusion of these counterparts can be analyzed for the SSD task. However, such combination of magnitude and phase information can be employed for the other successful feature sets (which utilizes only magnitude or phase information) in the literature.
- Furthermore, improved approaches of IF estimation can be utilized to further enhance the performance of the proposed CFCCIF-ESA feature set. For example, IFs can be estimated using fractional Hilbert transform and eigenvector theory as proposed in [372]. This approach defines smooth IF by removing the jump discontinuities caused by unwrapping of the instantaneous phase.
- If anti-spoofing will be implemented for banking applications, then transmission of the speech signal is transmitted over the long distance using the wired or wireless transmission line. It utilizes the speech coding and other signal transformations during this process. If SSD system is implemented after this transmission process, then the coding mechanism and transformations should be analyzed for the anti-spoofing task. The statistically meaningful dataset can be developed in this regard.
- For the experiments on environment-dependent scenario, each environment was partitioned into two disjoint and speaker-independent sets of roughly

the same size. The environment-wise statistics of the ReMASC dataset are shown in Table 3.14. The results obtained using CQCC, TECC, and ETECC feature sets with GMM classifier are reported in Table 4.18. Particularly for this scenario, we reported the results with the application of the CMVN for each utterance, as it has shown significant improvement. The analysis for the application of the CMN/CMVN techniques on environment-dependent *vs.* -independent scenario is discussed in Chapter 6 (Section 6.2). This needs further investigation and remains an open research problem. Notably, TECC performs better than the ETECC feature set for all the environments.

7.3 Open Research Problems

- It can be observed that even though the formal research in the anti-spoofing field started quarter-to-one decade before, however, still today there is no known statistically meaningful corpora for identical twins or professional impersonation; indicating the challenge associated with development of speech corpora for these two spoofs. Hence, the risk associated w.r.t. these two spoofing attacks for ASV system is unknown and hence, it continues to be a serious limitation in the anti-spoofing research field. This issue can be alleviated in future by developing the statistically meaningful corpora for identical twins or professional impersonation.
- It can be also observed that TECC feature set performs better than the ETECC feature set in Environment-D for environment-independent case. It might be due to noise suppression capability of the TEO especially for the vehicle noise (as originally reported for noise robust speech recognition in car [213]), whereas establishing noise suppression capability of ETEO remains an open research question for the future study.
- In this thesis, various feature sets are developed for individual spoofing attacks. CQCC and various variants of the CQT-based feature sets are proved to perform good for various spoofing attacks. However, various other feature sets performs better than CQCC features for various spoofing attacks. The development of a feature set that can perform better for all the spoofing attacks still remains an open research problem.
- It is also observed that TECC feature set performs better than ETECC feature set for ReMASC dataset in Env-D for environment-independent case. It might be due to noise suppression capability of the TEO especially for the

vehicle noise (as originally reported for noise robust speech recognition in car [213]), whereas noise suppression capability of ETEO remains an open research question for the future study.

- The replayed version of the genuine speech signal includes additional components which are impulse responses of playback device, playback environment, recording device, and recording environment [72]. Individual contribution of these additional components in the replay mechanism can be analyzed. It may help to develop the feature set based on the contribution of the individual components.
- For speech technology applications, perceptually-motivated feature representations are utilized, where the frequency bins are logarithmically or geometrically separated. However, in SSD task, it is observed that the feature sets (e.g., LFCC) which are linearly separated are giving relatively better performance. The analysis of this fact remains an open research question for the future study.
- The addition of the pop noise in the spoof speech utterance can be easily performed. However, the proposed approach did not address the issue of artificially added pop noise in spoof speech signal, which can be easily added at the arbitrary locations in the utterance. Thus, the VLD system can be further modified to detect the pop noise at pop noise-specific phonemes and improve the security for the ASV system, which remains an open research problem.
- The concept of the TEO is derived by estimating the total energy at a instant in SHM due to spring-mass system. The concept of CTEO can be derived using analogy from physical system.

Appendix A. Heisenberg's Uncertainty Principle in Signal Processing Framework

Theorem (Heisenberg's Uncertainty Principle): The temporal variance, σ_t^2 and the frequency variance, σ_ω^2 of a window $f(t) \in L^2(\mathbb{R})$ and having unit norm and fast decay satisfy,

$$\sigma_t^2 \cdot \sigma_\omega^2 \geq \frac{1}{4}. \quad (\text{A.1})$$

This inequality becomes *equality* if and only if $f(t)$ is a Gaussian window function.

Proof: This proof assumes fast decay of the window function $f(t) \in L^2(\mathbb{R})$ [373], however, this theorem is valid for any $f(t), t \cdot f(t),$ and $f'(t) \in L^2(\mathbb{R})$ [200]. Let us consider the integral I as,

$$I = \int_{t \in \mathbb{R}} (t \cdot f(t))(f'(t))dt = \langle tf(t), f'(t) \rangle. \quad (\text{A.2})$$

Using Cauchy-Schwartz inequality for vectors a and b , we have,

$$|\langle a, b \rangle| \leq \|a\| \cdot \|b\|. \quad (\text{A.3})$$

Hence,

$$|\langle tf(t), f'(t) \rangle| \leq \|tf(t)\| \cdot \|f'(t)\|, \quad (\text{A.4})$$

$$\left| \int_{t \in \mathbb{R}} tf(t)f'(t)dt \right| \leq \left[\int_{-\infty}^{+\infty} |tf(t)|^2 dt \right]^{\frac{1}{2}} \times \left[\int_{-\infty}^{+\infty} |f'(t)|^2 dt \right]^{\frac{1}{2}}. \quad (\text{A.5})$$

Since window $f(t)$ has unit norm, i.e., $\|f(t)\| = 1$, we have,

$$\int_{-\infty}^{+\infty} t^2 |f(t)|^2 dt = \sigma_t^2. \quad (\text{A.6})$$

Using Plancherel's theorem,

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\mathcal{F}(f(t))|^2 d\omega, \quad (\text{A.7})$$

where $\mathcal{F}(f(t))$ represents the Fourier transform of $f(t)$. Hence,

$$\int_{-\infty}^{+\infty} |f'(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\mathcal{F}(f'(t))|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \omega^2 |F(\omega)|^2 d\omega. \quad (\text{A.8})$$

Using the definition of σ_ω^2 in eq. (A.8), eq. (A.5) becomes,

$$|I|^2 \leq \sigma_t^2 \cdot \sigma_\omega^2. \quad (\text{A.9})$$

Using integration by parts on eq. (A.2), we get,

$$I = \left[\left(\frac{t}{2} \right) \int \frac{d}{dt} f^2(t) dt \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \left(\frac{d}{dt} \left(\frac{t}{2} \right) \cdot \int \frac{d}{dt} f^2(t) dt \right) dt \Big\} = \frac{-1}{2}. \quad (\text{A.10})$$

Hence, eq. (A.9) becomes,

$$\sigma_t^2 \cdot \sigma_\omega^2 \geq \frac{1}{4}. \quad (\text{A.11})$$

Cauchy-Schwartz's inequality becomes equality for collinear vectors, i.e., $b = -ka$.

$$\therefore f'(t) = -ktf(t), \quad (\text{A.12})$$

where k is a scalar such that $k > 0$.

$$\frac{f'(t)}{f(t)} = -kt. \quad (\text{A.13})$$

Solving the differential in eq. (A.13), we obtain,

$$\int \frac{df(t)}{f(t)} = \int -ktdt, \quad (\text{A.14})$$

$$\log_e f(t) = -kt^2, \quad (\text{A.15})$$

$$f(t) = e^{-kt^2}. \quad (\text{A.16})$$

It can be observed that eq. (A.16) represents the Gaussian function. Thus, this result proves that the lower bound on Heisenberg's box (i.e., $\sigma_t^2 \cdot \sigma_\omega^2$) is achieved for Gaussian window function. In particular, optimal localization is achieved by the family of Gabor atoms.

Appendix B. IF Estimation using Hilbert Transform

The analytic signal representation of the real signal $x_i(t)$ is given by:

$$x_{a_i}(t) = x_i(t) + jx_{h_i}(t), \quad (\text{B.1})$$

where,

$$x_{h_i}(t) = p.v. \int_{\tau=-\infty}^{\tau=+\infty} x_i(\tau) \left[\frac{1}{\pi(t-\tau)} \right] d\tau. \quad (\text{B.2})$$

when τ crosses t , the integrand in Eq. (B.2) becomes infinity and hence, the integral diverges. Such integrals are called as *singular* integrals in the mathematical literature [257]. To address this issue, we work in Fourier-domain by invoking the convolution theorem and *duality* property of CTFT, i.e.,

$$\mathcal{F}[x_{h_i}(t)] = \mathcal{F}[x_i(t)] \cdot \mathcal{F}\left[\frac{1}{\pi t}\right], \quad (\text{B.3})$$

$$X_{h_i}(\omega) = \begin{cases} +jX_i(\omega), & \text{for } \omega < 0, \\ -jX_i(\omega), & \text{for } \omega > 0. \end{cases} \quad (\text{B.4})$$

Thus, we need to use Fourier transform (and its inverse) to generate an analytic signal. In particular,

$$X_i(\omega) = \mathcal{F}[x_i(t)] = \int_{-\infty}^{+\infty} x_i(t)e^{-j\omega t} dt, \quad (\text{B.5})$$

Eq. (B.5) represents a global average of signal $x_i(t)$ with infinite duration sinusoidal waves, i.e., $\{e^{j\omega t}\}_{t \in \mathbb{R}}$. Thus, by using Heisenberg's uncertainty principle in signal processing framework [200], if we try to improve time resolution of $x(t)$ for IF estimation (say by multiplying $x(t)$ with a short-time window), we loose frequency resolution and vice-versa.

The instantaneous amplitude and instantaneous phase of the analytic signal $x_{a_i}(t)$ is given by:

$$|x_{a_i}(t)| = \sqrt{x_i^2(t) + x_{h_i}^2(t)}, \quad (\text{B.6})$$

$$\phi_i(t) = \tan^{-1}\left(\frac{x_{h_i}(t)}{x_i(t)}\right). \quad (\text{B.7})$$

IF is nothing but the derivative of the unwrapped instantaneous phase, $\phi_i(t)$, and it is expressed as:

$$IF = \frac{d}{dt}(\phi_i(t)). \quad (\text{B.8})$$

Due to the periodicity property of *arc-tangent* function, phase given by eq. (B.7) gets unwrapped to $-\pi$ to $+\pi$ or 0 to 2π interval and thus, creates discontinuity in the phase function, which makes it difficult for computation of derivative to get IF and thus, computationally complex task of phase unwrapping is required [374,375].

Appendix C. IF Estimation using ESA

In [31], three Discrete-time Energy Separation Algorithm (DESA) algorithms are mentioned, namely, DESA-1a, DESA-1, and DESA-2. In DESA-1a, '1' implies to derivative approximation in TEO with single sample difference, and a implies to the asymmetric difference. In DESA-1, the derivative operation is supposed to be symmetrized by averaging the two opposite asymmetric derivatives, namely, forward and backward differences. However, DESA-2 utilizes the symmetric 2-point sample difference to approximate the derivative operation. In this thesis, DESA-1a is utilized for energy separation [31, 183].

Let us consider a discrete-time AM-FM signal $y(n) = a(n) \cos(\phi(n))$, whose instantaneous frequency signal $\omega_i(n)$ is a finite sum of cosines. Its backward difference is given as:

$$\begin{aligned} s(n) &= y(n) - y(n-1), \\ &= a(n)c(n) + [a(n) - a(n-1)] \cos(\phi(n-1)), \\ &= D(n) + E(n), \end{aligned} \quad (\text{C.1})$$

where,

$$D(n) = a(n)c(n), \quad (\text{C.2})$$

$$E(n) = a(n)c(n) + [a(n) - a(n-1)] \cos(\phi(n-1)). \quad (\text{C.3})$$

Furthermore,

$$\begin{aligned} c(n) &= \cos(\phi(n)) - \cos(\phi(n-1)), \\ &= 2 \sin\left(\frac{\phi(n) + \phi(n-1)}{2}\right) \cdot \sin\left(\frac{\phi(n-1) - \phi(n)}{2}\right). \end{aligned} \quad (\text{C.4})$$

Using general approximations results for $\phi(n)$:

$$\phi(k) + \phi(m) \approx 2\phi\left(\frac{k+m}{2}\right) \quad \text{if } \omega_f |k-m| \ll 2. \quad (\text{C.5})$$

$$\phi(k) - \phi(m) \approx (k - m)\omega_i \frac{k + m}{2} \quad \text{if } \omega_f |k - m| \ll 2. \quad (\text{C.6})$$

If $\omega_f \ll 1$, we obtain from eq. (C.4):

$$c(n) \approx -2 \sin(\omega_i(n - 0.5)/2) \sin(\phi(n - 0.5)). \quad (\text{C.7})$$

Furthermore, according to Lemma 2 in [31], the order of magnitude of E and D in eq. (C.1) are:

$$\begin{aligned} D_{max} &\approx 2 \sin(\omega_i/2)_{max} a_{max}, \\ E_{max} &\approx 2 \sin(\omega_a/2) a_{max}. \end{aligned} \quad (\text{C.8})$$

If $a(n)$ in bandlimited, then the order of magnitude of D is much larger than that of E . Thus, ignoring E :

$$y(n) \approx -2a(n) \sin(\omega_i(n - 0.5)/2) \sin(\phi(n - 0.5)). \quad (\text{C.9})$$

Considering the first order approximation for standard series expansions for $\sin(\cdot)$ and $\cos(\cdot)$ on bandlimited signal:

$$\psi(s(n)) \approx 4a^2(n) \sin^2(\omega_i(n - 0.5)/2) \sin^2(\omega_i(n - 0.5)). \quad (\text{C.10})$$

Ignoring the half-sample shift and applying concept of TEO to discrete time signal, i.e., $\psi(y(n)) \approx a^2(n)\omega_i^2(n)$, we obtain:

$$|a(n)| \approx \sqrt{\frac{2\psi\{y(n)\}}{1 - \left(1 - \frac{\psi\{y(n) - y(n-1)\}}{2\psi\{y(n)\}}\right)'}} \quad (\text{C.11})$$

$$\omega_{if}(n) = \arccos \left[1 - \frac{\psi\{y(n) - y(n-1)\}}{2\psi\{y(n)\}} \right]. \quad (\text{C.12})$$

Appendix D. Cauchy-Schwarz Inequality for Multichannel Noise Power

In this Appendix, we follow the original work reported in [186] for more detailed and elaborate discussion here *w.r.t.* proposed CTECC_{max} framework for replay SSD on VAs. For simplicity of analysis, we make a basic assumption that the speech signal $s(t)$ is well approximated by an AM-FM signal, i.e., $s(t) = a(t) \cos(\phi(t))$ with both time-varying amplitude $a(t)$ and time-varying instantaneous frequency $\omega_i(t) = d\phi(t)/dt$. Such an assumption is valid because existence of AM-FM in speech resonance and also the place theory of hearing coupled with mathematical modelling of cochlea [376]. To that effect, Teager energy of $s(t)$ will be given by $\psi\{s(t)\} \approx a^2(t)\omega_i^2(t)$. Under this assumption and using the eq (4.52), the subband (bandpass) signal (in Gabor filterbank of CTECC) can be approximated as the output of j^{th} LTI filter as [186, 233]:

$$\hat{s}_j(t) = a(t)|G_j(\omega_i(t))| \cos(\phi(t) + \angle G_j(\omega_i(t))), \quad (\text{D.1})$$

where $G_j(\omega) = |G_j(\omega(t))| \angle G_j(\omega(t))$ is a frequency response of j^{th} Gabor filter in the filterbank. In particular, $|G_j(\omega)|$ and $\angle G_j(\omega_i(t))$ are called as *gain* and *phase shift* of LTI filter, $g_i(t)$ [366]. Thus, the TEO of j^{th} subband filtered signal $s_j(t)$ will be equal to

$$\psi\{s_j(t)\} = (a(t)|G_j(\omega_i(t))|)^2 \omega_i^2(t), \quad (\text{D.2})$$

$$\therefore \psi\{s_j(t)\} = a^2(t)|G_j(\omega_i(t))|^2 \omega_i^2(t). \quad (\text{D.3})$$

Here, we focus on second term of eq. (4.55), i.e., $E\{\Psi_{cr}[n_{p_j}(t), n_{q_j}(t)]\}$. The noise processes $n_{p_j}(t)$ and $n_{q_j}(t)$ have cross-power spectral density $\Phi(\omega_{pq}) = \mathcal{F}\{R_{pq}(\tau)\}$, where $\mathcal{F}\{\cdot\}$ denotes Fourier transform operator. Thus, the j^{th} subband noise process will have cross-power spectral density [228],

$$\Phi_{(pq)j}(\omega) = |G_j(\omega)|^2 \cdot \Phi_{pq}(\omega). \quad (\text{D.4})$$

Further, because $n_p(t)$ and $n_q(t)$ are Wide Sense Stationary (WSS) Gaussian, their derivatives, i.e., the processes $\dot{n}_p(t)$ and $\dot{n}_q(t)$ are also WSS Gaussian, and their product $\dot{n}_p(t) \cdot \dot{n}_q(t)$ is statistically-independent of both $n_{p_j}(t)$ and $\dot{n}_{q_j}(t)$ [228]. Hence, CTEO between $n_{p_j}(t)$ and $n_{q_j}(t)$ is given by:

$$\Psi_{cr}\{n_{p_j}(t), n_{q_j}(t)\} = \dot{n}_{p_j}(t)\dot{n}_{q_j}(t) - n_{p_j}(t)\ddot{n}_{q_j}(t), \quad (\text{D.5})$$

and it is the sum of two independent random processes. To estimate the mean of this, we need to estimate the following:

$$E[\dot{n}_{p_j}(t)\dot{n}_{q_j}(t)] = -R_{(pq)_j}^{(2)}(0), \quad (\text{D.6})$$

$$E[n_{p_j}(t)\ddot{n}_{q_j}(t)] = R_{(pq)_j}^{(2)}(0). \quad (\text{D.7})$$

Using Wiener-Khinchin theorem [377], we have the autocorrelation function $R_{(p)}(\tau)$ given by,

$$R_p(\tau) = \mathcal{F}^{-1}\{\Phi_p(\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi_p(\omega) \cdot e^{j\omega\tau} d\omega. \quad (\text{D.8})$$

Similarly,

$$R_q(\tau) = \mathcal{F}^{-1}\{\Phi_q(\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi_q(\omega) \cdot e^{j\omega\tau} d\omega. \quad (\text{D.9})$$

Hence, cross-correlation function $R_{(pq)_j}(\tau)$ corresponding to the j^{th} subband is given as,

$$R_{(pq)_j}(\tau) = \mathcal{F}^{-1}\{\Phi_{(pq)_j}(\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi_{(pq)_j}(\omega) \cdot e^{j\omega\tau} d\omega. \quad (\text{D.10})$$

Differentiating eq. (D.10) w.r.t. τ two times under integral sign, we get,

$$R_{(pq)_j}^{(2)}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi_{(pq)_j}(\omega) \cdot \frac{d^2(e^{j\omega\tau})}{d\tau^2} d\omega, \quad (\text{D.11})$$

$$R_{(pq)_j}^{(2)}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (j\omega)^2 \Phi_{(pq)_j}(\omega) \cdot e^{j\omega\tau} d\omega. \quad (\text{D.12})$$

At zeroth lag, $\tau = 0$, we have,

$$R_{(pq)_j}^{(2)}(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (j\omega)^2 \Phi_{(pq)_j}(\omega) d\omega. \quad (\text{D.13})$$

Using eq. (D.4) in eq. (D.13), we get

$$R_{(pq)_j}^{(2)}(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (j\omega)^2 |G_j(\omega)|^2 \Phi_{(pq)}(\omega) d\omega, \quad (\text{D.14})$$

which can be approximated as [233]:

$$R_{(pq)_j}^{(2k)}(0) = \hat{R}_{(pq)_j}^{(2k)}(\omega_i(t)), \quad (\text{D.15})$$

where

$$\hat{R}_{(pq)_j}^{(2k)}(\omega_i(t)) = (-1)^k \omega_i^{2k}(t) |G_j(\omega_i(t))|^2 \Gamma_{(pq)_j}, \quad (\text{D.16})$$

with

$$\Gamma_{(pq)_j} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|G_j(\omega)|^2}{|G_j(\omega_c)|^2} \Phi_{(pq)}(\omega) d\omega, \quad (\text{D.17})$$

representing concentration of noise power within the passband of j^{th} subband filter $g_j(t)$, and we represented center frequency of j^{th} filter. Using eq. (4.55), eq. (D.3), eq. (D.5), eq. (D.6), eq. (D.7), eq. (D.16), we get the mean of $\Psi_{cr}\{x_{p_j}(t), x_{q_j}(t)\}$ as:

$$E\{\Psi_{cr}[x_{p_j}(t), x_{q_j}(t)]\} = E\{\Psi_{cr}[s_j(t)]\} + 2\omega_i^2(t) |G_j(\omega_i(t))|^2 \tau_{(pq)_j} \quad (\text{D.18})$$

The second term of RHS of eq. (D.18) corresponds to error term. Using Cauchy-Schwartz inequality as in [228]:

$$\left| \int_{-\infty}^{+\infty} \Phi_{(pq)_j}(\omega) d\omega \right|^2 \leq \left[\int_{-\infty}^{+\infty} \Phi_{p_j}(\omega) d\omega \right] \left[\int_{-\infty}^{+\infty} \Phi_{q_j}(\omega) d\omega \right], \quad (\text{D.19})$$

which gives,

$$\left| \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|G_j(\omega)|^2}{|G_j(\omega_c)|^2} \Phi_{(pq)_j}(\omega) d\omega \right|^2 \leq \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|G_j(\omega)|^2}{|G_j(\omega_c)|^2} \Phi_{p_j}(\omega) d\omega \right] \cdot \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|G_j(\omega)|^2}{|G_j(\omega_c)|^2} \Phi_{q_j}(\omega) d\omega \right], \quad (\text{D.20})$$

$$\left| R_{(pq)_j}(\tau) \right|_{\tau=0}^2 \leq R_{p_j}(\tau) \Big|_{\tau=0} \cdot R_{q_j}(\tau) \Big|_{\tau=0}, \quad (\text{D.21})$$

$$\left| R_{(pq)_j}(0) \right|^2 \leq R_{p_j}(0) \cdot R_{q_j}(0), \quad (\text{D.22})$$

which leads to the following inequality to analyze noise power in j^{th} subband using eq. (4.57),

$$|\Gamma_{(pq)_j}|^2 \leq \Gamma_{p_j} \Gamma_{q_j}. \quad (\text{D.23})$$

The inequality (D.23) is useful to analyze the efficiency of proposed CTECC feature extraction framework for replay SSD task on VAs. In particular, it can be observed from inequality (D.23) that proposed idea of maximum energy distortions in CTECC_{max} framework is indeed the optimal solution *w.r.t.* concentration of the noise power in j^{th} subband *via* CTEO framework than its individual channel (i.e., Γ_{p_j} or Γ_{q_j}) counterpart.

Bibliography

- [1] A. T. Patil, H. A. Patil, and K. Khorria, "Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection," *Computer Speech & Language*, vol. 72, p. 101301, 2022.
- [2] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, Dresden, Germany, 6-11 Sept., 2015, pp. 2037–2041.
- [3] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 26 - 29 June, 2018, pp. 296–303.
- [4] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.
- [5] 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, Sept. 15-19, 2019, pp. 1008–1012.
- [6] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Arlington, USA, Sept. 8-11, 2015, pp. 1–6.
- [7] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. S. Mello, R. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition," in *8th International Conference on Biometrics Theory, Applications, and Systems (BTAS)*, Niagara Falls, Buffalo, USA, Sept. 6-9, 2016, pp. 1–6.
- [8] K. Akimoto, S. P. Liew, S. Mishima, R. Mizushima, and K. A. Lee, "POCO: A voice spoofing and liveness detection corpus based on pop noise," in *INTERSPEECH*, Shanghai, China, October 25-29, 2020, pp. 1081–1085.
- [9] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 2355–2359.

- [10] A. T. Patil, R. Acharya, H. A. Patil, and R. C. Guido, "Improving the potential of enhanced teager energy cepstral coefficients (ETECC) for replay attack detection," *Computer Speech & Language*, vol. 72, p. 101281, 2022.
- [11] R. Acharya, H. Kotta, A. T. Patil, and H. A. Patil, "Cross-teager energy cepstral coefficients for replay spoof detection on voice assistants," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June 6-11, 2021, pp. 6364–6368.
- [12] Y. Gong, J. Yang, and C. Poellabauer, "Detecting replay attacks using multi-channel audio: A neural network-based method," *IEEE Signal Processing Letters*, vol. 27, pp. 920–924, 2020.
- [13] A. T. Patil, A. Therattil, and H. A. Patil, "On significance of cross-teager energy cepstral coefficients for replay spoof detection on voice assistants," *under review in Computer Speech & Language*, 2022.
- [14] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America (JASA)*, vol. 89, no. 1, pp. 425–434, 1991.
- [15] K. Khorra, A. T. Patil, and H. A. Patil, "On significance of constant-q transform for pop noise detection," *Computer Speech & Language*, p. 101421, 2022.
- [16] C. F. Eyring, "Reverberation time in "dead" rooms," *The Journal of the Acoustical Society of America (JASA)*, vol. 1, no. 2A, pp. 217–241, 1930.
- [17] 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.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, June 26th - July 1st, 2016, pp. 770–778.
- [19] A. T. Patil, H. Kotta, R. Acharya, and H. A. Patil, "Spectral root features for replay spoof detection in voice assistants," in *SPECOM, Petersburg, Russia*. Springer, Sept. 27-30, 2021, pp. 504–515.
- [20] P. Tapkir and H. A. Patil, "Novel empirical mode decomposition cepstral features for replay spoof detection," in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 721–725.
- [21] A. T. Patil and H. A. Patil, "Significance of CMVN for replay spoof detection," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, Dec. 7-10, 2020, pp. 532–537.
- [22] P. Chodingala, S. Chaturvedi, A. T. Patil, and H. A. Patil, "Robustness of DAS beamformer over MVDR for replay attack detection on voice assistants," in *accepted in International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, July 11-15, 2022.
- [23] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH*, Brisbane, Australia, September 22-26, 2008, pp. 1741–1744.

- [24] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, 2021.
- [25] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi, "Classifying infant cry patterns by the genetic selection of a fuzzy model," *Biomedical Signal Processing and Control*, vol. 17, pp. 38–46, 2015.
- [26] H. A. Patil, A. T. Patil, and A. Kachhi, "Constant Q cepstral coefficients for classification of normal vs. pathological infant cry," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 22-27, 2022, pp. 7392–7396.
- [27] A. T. Patil, A. Kachhi, and H. A. Patil, "Subband teager energy representations for infant cry analysis and classification," in *accepted in European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, August 29 - Sept. 2, 2022.
- [28] 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.
- [29] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 618–631, 2016.
- [30] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi *et al.*, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.
- [31] P. Maragos, J. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.
- [32] R. Acharya, H. A. Patil, and H. Kotta, "Novel enhanced Teager energy based cepstral coefficients for replay spoof detection," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, December 14-18, 2019, pp. 342–349.
- [33] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.
- [34] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH, Dresden, Germany*, Sept. 6-10, 2015, pp. 239–243.
- [35] K. Khoría, A. T. Patil, and H. A. Patil, "Significance of constant-Q transform for voice liveness detection," in *European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, August 23-27, 2021, pp. 126–130.
- [36] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America (JASA)*, vol. 65, no. 4, pp. 943–950, 1979.

- [37] A. Meyer, D. Döbler, J. Hambrecht, and M. Matern, "Acoustic mapping on three-dimensional models," in *Proceedings of the 12th International Conference on Computer Systems and Technologies*, Vienna, Austria, 2011, pp. 216–220.
- [38] H. Bořil and P. Pollák, "Direct time domain fundamental frequency estimation of speech in noisy conditions," in *12th European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, September 1–5, 2004, pp. 1003–1006.
- [39] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with parkinson's disease." in *INTERSPEECH, Stockholm, Sweden*, September 14-18, 2017, pp. 314–318.
- [40] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *INTERSPEECH*, Lisbon, Portugal, Sept. 4-8, 2005, pp. 3013–3016.
- [41] M. R. Kamble and H. A. Patil, "Detection of replay spoof speech using teager energy feature cues," *Computer Speech & Language*, vol. 65, p. 101140, 2021.
- [42] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 1st Edition, Pearson Education India, 2015.
- [43] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, Orlando, FL, USA, May 13-17, 2002, pp. IV–4072–IV–4075.
- [44] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.
- [45] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.
- [46] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.
- [47] 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.
- [48] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing*, 2nd ed. Springer, 2018.
- [49] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.
- [50] 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.
- [51] 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)*, Honolulu, Hawaii, USA, Nov. 12-15, 2018, pp. 1047–1053.

- [52] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 25-30, 2012, pp. 4401–4404.
- [53] J. Eargle, "In-line, planar loudspeakers, and arrays," in *Loudspeaker Handbook*. Springer, 2003, pp. 133–160.
- [54] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 27–31.
- [55] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.
- [56] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.
- [57] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *proceedings of the international conference of biometrics special interest group (BIOSIG)*, Sept. 6-7, 2012, pp. 1–7.
- [58] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial" gummy" fingers on fingerprint systems," in *Optical Security and Counterfeit Deterrence Techniques IV*, vol. 4677, 2002, pp. 275–289.
- [59] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*, March 29 - April 1, 2012, pp. 26–31.
- [60] *Spoofing and Countermeasures for Automatic Speaker Verification, Special Sessions in INTERSPEECH-2013*. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2013/interspeech_2013.pdf
- [61] 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.
- [62] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector," in *Speaker Odyssey*, June 21-24, 2016, pp. 259–263.
- [63] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proce. of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, TX, USA, Oct. 30 - Nov. 3, 2017, pp. 103–117.
- [64] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, "Towards protecting cyber-physical and iot systems from single-and multi-order voice spoofing attacks," *Applied Acoustics*, vol. 183, p. 108283, 2021.

- [65] 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, Sept. 6-10, 2015, pp. 2062–2066.
- [66] J. Lim, "Spectral root homomorphic deconvolution system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.
- [67] 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.
- [68] Y. Stylianou, "Voice transformation: A survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 19-24, 2009, pp. 3585–3588.
- [69] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [70] 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.
- [71] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, China, October 20-24, 2004, pp. 145–148.
- [72] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, September 10-12, 2014, pp. 1–6.
- [73] A. Paul, R. K. Das, R. Sinha, and S. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *International Conference on Signal Processing and Communications (SPCOM)*, IISc, Bengaluru, India, June 12-15 2016, pp. 1–5.
- [74] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *INTERSPEECH*, Brighton, United Kingdom, Sept. 6-10, 2009, pp. 1559–1562.
- [75] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVSpooF 2015 challenge," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 20-25, 2016, pp. 5475–5479.
- [76] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2087–2091.
- [77] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2042–2046.

- [78] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2092–2096.
- [79] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2082–2086.
- [80] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2057–2061.
- [81] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2072–2076.
- [82] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *INTERSPEECH*, Dresden, Germany, Sept. 6-10, 2015, pp. 2077–2081.
- [83] 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, Bilbao, Spain, June 21-24, 2016, pp. 249–252.
- [84] 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, 2016.
- [85] D. Paul, M. Pal, and G. Saha, "Spectral features for synthetic speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 605–617, 2017.
- [86] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (SoE) for spoofed speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 20-25, 2016, pp. 5105–5109.
- [87] T. B. Patel and H. A. Patil, "Significance of source–filter interaction for classification of natural vs. spoofed speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 644–659, 2017.
- [88] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 660–670, 2017.
- [89] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Computer Speech & Language*, vol. 48, pp. 31–50, 2018.
- [90] 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, Sept. 6-10, 2015.

- [91] 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.
- [92] 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, Sept. 6-10, 2015, pp. 2067–2071.
- [93] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 651–655.
- [94] 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, Sept. 6-10, 2015, pp. 2097–2101.
- [95] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 4, pp. 2160–2170, 2019.
- [96] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2373–2384, 2019.
- [97] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention." in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 681–685.
- [98] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep siamese architecture based replay detection for secure voice biometric." in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 671–675.
- [99] M. Saranya and H. A. Murthy, "Decision-level feature switching as a paradigm for replay attack detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 686–690.
- [100] L. Huang and C.-M. Pun, "Audio replay spoof attack detection using segment-based hybrid feature and densenet-LSTM network," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17,, pp. 2567–2571.
- [101] B. Wickramasinghe, E. Ambikairajah, J. Epps, V. Sethu, and H. Li, "Auditory inspired spatial differentiation for replay spoofing attack detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17,, pp. 6011–6015.
- [102] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17,, pp. 6201–6205.
- [103] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 666–670.

- [104] 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.
- [105] A. T. Patil, A. Rajul, P. Sai, and H. A. Patil, "Energy separation-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 2898–2902.
- [106] M. R. Kamble and H. A. Patil, "Analysis of reverberation via Teager energy features for replay spoof speech detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 2607–2611.
- [107] 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 2017, pp. 7–11.
- [108] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1013–1017.
- [109] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble Models for Spoofing Detection in Automatic Speaker Verification," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1018–1022.
- [110] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU replay detection system for the asvspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1023–1027.
- [111] R. Bialobrzeski, M. Kosmider, M. Matuszewski, M. Plata, and A. Rakowski, "Robust Bayesian and Light Neural Networks for Voice Spoofing Detection," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1028–1032.
- [112] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1033–1037.
- [113] K. R. Alluri and A. K. Vuppala, "IIIT-H Spoofing Countermeasures for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2019," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1043–1047.
- [114] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on asvspoof 2019," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, Dec. 14-18, 2019, pp. 1018–1025.
- [115] R. K. Das, J. Yang, and H. Li, "Long Range Acoustic Features for Spoofed Speech Detection," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1058–1062.
- [116] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1068–1072.

- [117] H. Zeinali, T. Stafylakis, G. Athanasopoulou, J. Rohdin, I. Gkinis, L. Burget, and J. Cernocky, "Detecting Spoofing Attacks Using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1073–1077.
- [118] R. Li, M. Zhao, Z. Li, L. Li, and Q. Hong, "Anti-Spoofing Speaker Verification System with Multi-Feature Integration and Multi-Task Learning," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1048–1052.
- [119] J. Williams and J. Rownicka, "Speech Replay Detection with x-Vector Attack Embeddings and Spectral Features," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1053–1057.
- [120] M. G. Kumar, S. R. Kumar, M. Saranya, B. Bharathi, and H. A. Murthy, "Spoof detection using time-delay shallow neural network and feature switching," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, Dec. 14-18, 2019, pp. 1011–1017.
- [121] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, Dec. 14-18, 2019, pp. 1003–1010.
- [122] S.-Y. Chang, K.-C. Wu, and C.-P. Chen, "Transfer-Representation Learning for Detecting Spoofing Attacks with Converted and Synthesized Speech in Automatic Speaker Verification System," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1063–1067.
- [123] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1078–1082.
- [124] J. weon Jung, H. jin Shim, H.-S. Heo, and H.-J. Yu, "Replay Attack Detection with Complementary High-Resolution Information Using End-to-End DNN for the ASVspoof 2019 Challenge," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 1083–1087.
- [125] B. Chettri, T. Kinnunen, and E. Benetos, "Deep generative variational autoencoding for replay spoof detection in automatic speaker verification," *Computer Speech & Language*, vol. 63, p. 101092, 2020.
- [126] H. Wu, S. Liu, H. Meng, and H.-y. Lee, "Defense against adversarial attacks on spoofing countermeasures of asv," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 4-8, 2020, pp. 6564–6568.
- [127] S.-H. Yoon and H.-J. Yu, "Multiple points input for convolutional neural networks in replay attack detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 4-8, 2020, pp. 6444–6448.
- [128] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 4-8, 2020, pp. 6589–6593.

- [129] J. Monteiro, J. Alam, and T. H. Falk, "An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 4-8, 2020, pp. 6599–6603.
- [130] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, "The SJTU robust anti-spoofing system for the ASVspooF 2019 challenge." in *INTERSPEECH, Graz, Austria*, Sept. 15-19, 2019, pp. 1038–1042.
- [131] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 1–8.
- [132] L. Zhang, X. Wang, E. Cooper, and J. Yamagishi, "Multi-task Learning in Utterance-level and Segmental-level Spoof Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 9–15.
- [133] Y. Lei, X. Huo, Y. Jiao, and Y. K. Li, "Deep Metric Learning for Replay Attack Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 42–46.
- [134] Z. Benhafid, S. A. Selouani, M. S. Yakoub, and A. Amrouche, "LARIHS ASSERT Reassessment for Logical Access ASVspooF 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 94–99.
- [135] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The DKU-CMRI System for the ASVspooF 2021 Challenge: Vocoder based Replay Channel Response Estimation," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 16–21.
- [136] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 22–28.
- [137] R. K. Das, "Known-unknown Data Augmentation Strategies for Detection of Logical Access, Physical Access and Speech Deepfake Attacks: ASVspooF 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 29–36.
- [138] S. Yoon and H.-J. Yu, "Multiple-Point Input and Time-Inverted Speech Signal for The ASVspooF 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 37–41.
- [139] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is Silver, Silence is Golden: What do ASVspooF-trained Models Really Learn?" in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 2021, pp. 55–60.
- [140] J. Cáceres, R. Font, T. Grau, and J. Molina, "The Biometric Vox System for the ASVspooF 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 2021, pp. 68–74.

- [141] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 75–82.
- [142] W. H. Kang, J. Alam, and A. Fathan, "Investigation on activation functions for robust end-to-end spoofing attack detection system," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 83–88.
- [143] W. H. Kang, J. Alam, and A. Fathan, "CRIM's System Description for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 100–106.
- [144] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 61–67.
- [145] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 16, 2021, pp. 89–93.
- [146] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection based on pop-noise detector with phoneme information for speaker verification," *The Journal of the Acoustical Society of America (JASA)*, vol. 140, no. 4, pp. 3060–3060, 2016.
- [147] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection using phoneme-based pop-noise detector for speaker verification," in *Odyssey 2018, The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, June 26-29,, pp. 233–239.
- [148] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE Conference on Computer Communications*, Paris, France, April 29 - May 2, 2019, pp. 2062–2070.
- [149] P. Gupta, S. Gupta, and H. A. Patil, "Voice liveness detection using bump wavelet with CNN," in *International Conference on Pattern Recognition and Machine Intelligence*, December 15 - 18, 2021.
- [150] S. Singh, K. Khorra, and H. A. Patil, "Modified group delay function using different spectral smoothing techniques for voice liveness detection," in *International Conference on Speech and Computer (SPECOM)*, Petersburg, Russia, Sept. 27-30, 2021, pp. 649–659.
- [151] A. T. Patil, K. Khorra, and H. A. Patil, "Voice liveness detection using constant-q transform-based features," in *European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, August 29 - Sept. 2 2022.
- [152] S. Gupta, K. Khorra, A. T. Patil, and H. A. Patil, "Deep convolutional neural network for voice liveness detection," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Tokyo, Japan, Dec. 14-17, 2021, pp. 775–779.

- [153] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik, "Voice spoofing detection corpus for single and multi-order audio replays," *Computer Speech & Language*, vol. 65, p. 101132, 2021.
- [154] A. Therattil, P. Gupta, P. K. Chodingala, and H. A. Patil, "Teager energy based-detection of one-point and two-point replay attacks: Towards cross-database generalization," in *Odyssey 2022, The Speaker and Language Recognition Workshop*, June 28 - July 1, 2022, pp. 47–54.
- [155] E. Bjornson, "Reproducible research: Best practices and potential misuse [perspectives]," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 106–123, 2019.
- [156] M. Barni and F. Perez-Gonzalez, "Pushing science into signal processing [my turn]," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 120–119, 2005.
- [157] J. Kovacevic, "How to encourage and publish reproducible research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, USA, April 15-20, 2007, pp. 1273–1276.
- [158] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, 2016.
- [159] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 37–47, 2009.
- [160] 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 2015, pp. 4440–4444.
- [161] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.
- [162] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.
- [163] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, California, USA, March 23-26, 1992, pp. 137–140.
- [164] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.
- [165] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The REDDOTS data collection for speaker recognition," in *INTERSPEECH*, Dresden, Germany, Sept. 4-5, 2015, pp. 2996–3000.

- [166] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning*, Last Accessed July 22, 2021. [Online]. Available: <https://datashare.ed.ac.uk/handle/10283/3443>
- [167] ICASSP-2021, *Multi-Speaker Multi-Style Voice Cloning Challenge (M2VoC)*, <http://challenge.ai.iqiyi.com/detail?raceId=5fb2688224954e0b48431fe0>, {Last Accessed: March 1, 2022}. [Online]. Available: <http://challenge.ai.iqiyi.com/detail?raceId=5fb2688224954e0b48431fe0>
- [168] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016, {Last Accessed: March 1, 2022}.
- [169] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [170] *ASVspoof-2021 Workshop, Satellite event in INTERSPEECH-2021*, 2021 (Last accessed May 25, 2021). [Online]. Available: <https://www.asvspoof.org/workshop>
- [171] G. Ekman, "Weber's law and related functions," *The Journal of Psychology*, vol. 47, no. 2, pp. 343–352, 1959.
- [172] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.
- [173] A.-O. Boudraa and F. Salzenstein, "Teager-kaiser energy methods for signal and image analysis: A review," *Digital Signal Processing*, vol. 78, pp. 338–375, 2018.
- [174] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [175] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.
- [176] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [177] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Paris, France, May 30 - June 2, 2010, pp. 253–256.
- [178] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.
- [179] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, Sept. 22-25, 1997, pp. 1895–1898.
- [180] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

- [181] N. Brümmner and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013 (Last accessed August 10, 2020).
- [182] 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, USA, April 3-6, 1990, pp. 381–384.
- [183] 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.
- [184] 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.
- [185] R. C. Guido, "Enhancing Teager energy operator based on a novel and appealing concept: Signal mass," *Journal of the Franklin Institute*, vol. 356, no. 4, pp. 2346–2352, 2019.
- [186] S. Lefkimmiatis, P. Maragos, and A. Katsamanis, "Multisensor multiband cross-energy tracking for feature extraction and recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, USA, March 30 - April 4, 2008, pp. 4741–4744.
- [187] 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.
- [188] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in asv systems," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 51–55.
- [189] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.
- [190] A. Georgogiannis and V. Digalakis, "Speech emotion recognition using non-linear Teager energy based features in noisy environments," in *20th European signal processing conference (EUSIPCO)*, Bucharest, Romania, August 27-31, 2012, pp. 2045–2049.
- [191] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.
- [192] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.
- [193] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech Production and Speech Modelling*, Springer, pp. 241–261, 1990.
- [194] P. Maragos, J.F. Kaiser and T.F. Quatieri, "On separating amplitude from frequency modulations using energy operators," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, California, USA, March 23-26, 1992, pp. 1–4.

- [195] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 641–645.
- [196] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Canada, April 14-17, 1991, pp. 421–424.
- [197] M. R. Kamble and H. A. Patil, "Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 646–650.
- [198] Q. Li, "An auditory-based transform for audio signal processing," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 18-21, 2009, pp. 181–184.
- [199] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, USA: SIAM, 1992.
- [200] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Elsevier, 1999.
- [201] G. Von Békésy and E. G. Wever, *Experiments in Hearing*. McGraw-Hill New York, 1960, vol. 8.
- [202] D. Gabor, "Theory of Communication. Part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.
- [203] B. C. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.
- [204] S. S. Stevens, "On the psychophysical law," *Psychological Review*, vol. 64, no. 3, p. 153, 1957.
- [205] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth Intl. conference on artificial intelligence and statistics*, Sardinia, Italy, 2010, pp. 249–256.
- [206] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014, {Last Accessed: Jan 30, 2021}.
- [207] R. C. Guido, "Paraconsistent feature engineering [Lecture Notes]," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 154–158, 2018.
- [208] P. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *Symposium on Hearing Theory*, IPO, Eindhoven, Holland, 1972, p. 58–69.
- [209] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, "Auditory-motivated gamma-tone wavelet transform," *Signal Processing*, vol. 94, pp. 608–619, 2014.
- [210] J. Hosken, "Ricker wavelets in their various guises," *First Break*, vol. 6, no. 1, 1988.
- [211] J. van Hout, L. Ferrer, D. Vergyri, N. Scheffer, Y. Lei, V. Mitra, and S. Wegmann, "Calibration and multiple system fusion for spoken term detection using linear logistic regression," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4-9, 2014, pp. 7138–7142.

- [212] R. O. Duda and P. E. Hart, *Pattern classification*. John Wiley & Sons, 2006.
- [213] 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.
- [214] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Security and Privacy Workshops (SPW)*, San Francisco, USA, May 2018, pp. 1–7.
- [215] J. F. Kaiser, "Some useful properties of teager's energy operators," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Minnesota, USA, April 27-30 1993, pp. 149–152.
- [216] P. Maragos and A. Potamianos, "Higher order differential energy operators," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 152–154, 1995.
- [217] R. Hamila, J. Astola, F. A. Cheikh, M. Gabbouj, and M. Renfors, "Teager energy and the ambiguity function," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 260–262, 1999.
- [218] J.-C. Cexus and A.-O. Boudraa, "Link between cross-Wigner distribution and cross-Teager energy operator," *Electronics Letters*, vol. 40, no. 12, pp. 778–780, 2004.
- [219] A.-O. Boudraa, J.-C. Cexus, M. Groussat, and P. Brunagel, "An energy-based similarity measure for time series," *EURASIP Journal on Advances in Signal Processing*, vol. vol. 2008, pp. pp. 1–8, 2007.
- [220] J. Montillet, "On a novel approach to decompose finite energy functions by energy operators and its application to the general wave equation," in *International Mathematical Forum*, vol. 48, 2010, pp. 2387–2400.
- [221] A.-O. Boudraa, S. Benramdane, J.-C. Cexus, and T. Chonavel, "Some useful properties of cross- ψ_b -energy operator," *AEU-International Journal of Electronics and Communications*, vol. vol. 63, no. 9, pp. pp. 728–735, 2009.
- [222] A.-O. Boudraa, J.-C. Cexus, and H. Zaidi, "Functional segmentation of dynamic nuclear images by cross- ψ_b -energy operator," *Computer Methods and Programs in Biomedicine*, vol. vol. 84, no. 2-3, pp. pp. 146–152, 2006.
- [223] Z. Saidi, A. Boudraa, J. Cexus, and S. Bourennane, "Time-delay estimation using cross- ψ_b -energy operator," *International Journal of Electrical and Computer Engineering*, vol. vol. 1, no. 9, pp. pp. 1440–1444, 2007.
- [224] F. Salzenstein, P. Montgomery, and A.-O. Boudraa, "Local frequency and envelope estimation by Teager-Kaiser energy operators in white-light scanning interferometry," *Optics Express*, vol. vol. 22, no. 15, pp. pp. 18 325–18 334, 2014.
- [225] A.-O. Boudraa, J.-C. Cexus, and K. Abed-Meraim, "Cross ψ_b -energy operator-based signal detection," *The Journal of the Acoustical Society of America (JASA)*, vol. 123, no. 6, pp. 4283–4289, 2008.

- [226] I. Rodomagoulakis and P. Maragos, "On the improvement of modulation features using multi-microphone energy tracking for robust distant speech recognition," in *25th European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, August 28 - Sept. 2, 2017, pp. 558–562.
- [227] R. Bhatia and C. Davis, "A Cauchy-Schwartz inequality for operators with applications," *Linear Algebra and Its Applications*, vol. 223, pp. 119–129, 1995.
- [228] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education 4th edition, 2002.
- [229] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.
- [230] S. Shamma and D. Klein, "The case of the missing pitch templates: how harmonic templates emerge in the early auditory system," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2631–2644, 2000.
- [231] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Communication*, vol. 53, no. 5, pp. 707–715, 2011.
- [232] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002.
- [233] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3245–3265, 1993.
- [234] 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.
- [235] P. Mowlae, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.
- [236] P. A. Tapkir, A. T. Patil, N. Shah, and H. A. Patil, "Novel spectral root cepstral features for replay spoof detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 12-15, 2018, pp. 1945–1950.
- [237] H. A. Patil and M. C. Madhavi, "Significance of magnitude and phase information via VTEO for humming-based biometrics," in *5th IAPR International Conference on Biometrics (ICB)*, New Delhi, India, March 29 - April 1, 2012, pp. 372–377.
- [238] H. A. Patil and M. C. Madhavi, "Combining evidences from magnitude and phase information using VTEO for person recognition using humming," *Computer Speech & Language*, vol. 52, pp. 225–256, 2018.
- [239] P. Mowlae, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," in *INTERSPEECH*, Singapore, Sept. 2014, pp. 1623–1627.

- [240] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1521–1532, 2015.
- [241] K. Vijayan, V. Kumar, and K. S. R. Murty, "Feature extraction from analytic phase of speech signals for speaker verification," in *INTERSPEECH*, Singapore, Sept. 14-18, 2014, pp. 1658–1662.
- [242] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining mfcc and phase information in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 14-19, 2010, pp. 4502–4505.
- [243] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2005.
- [244] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Seventh European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept. 3-7, 2001, pp. 759–762.
- [245] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.
- [246] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*. Springer, 2014, pp. 125–146.
- [247] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *A meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.
- [248] L. Cohen, *Time-Frequency Analysis*. Prentice-Hall, 1995, vol. 778.
- [249] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications*. MIT Press, Mass, 1949.
- [250] M. R. Schroeder and B. S. Atal, "Generalized short-time power spectra and auto-correlation functions," *The Journal of the Acoustical Society of America (JASA)*, vol. 34, no. 11, pp. 1679–1683, 1962.
- [251] G. Gambardella, "A contribution to the theory of short-time spectral analysis with nonuniform bandwidth filters," *IEEE Transactions on Circuit Theory*, vol. 18, no. 4, pp. 455–460, 1971.
- [252] R. A. Altes, "The Fourier–Mellin transform and mammalian hearing," *The Journal of the Acoustical Society of America (JASA)*, vol. 63, no. 1, pp. 174–183, 1978.
- [253] G. Gambardella, "The Mellin transforms and constant-q spectral analysis," *The Journal of the Acoustical Society of America (JASA)*, vol. 66, no. 3, pp. 913–915, 1979.

- [254] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Oklahoma, USA, April 10-12, 1978, pp. 375–378.
- [255] 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. 2–6.
- [256] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, Seim Reap, Cambodia, December 9-12, 2014, pp. 1–5.
- [257] A. V. Oppenheim, J. R. Buck, and R. W. Schaffer, *Discrete-Time Signal Processing*. Vol. 2. Upper Saddle River, NJ: Prentice Hall, 2001.
- [258] S. Maymon and A. V. Oppenheim, "Sinc interpolation of nonuniform samples," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4745–4758, 2011.
- [259] J. G. Proakis, *Digital Signal Processing: Principles Algorithms and Applications*. Pearson Education India, 2001.
- [260] D. Campbell, K. Palomaki, and G. Brown, "A matlab simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.
- [261] E. Vincent and D. R. Campbell, "Roomsimove," *GNU Public License*, http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip, vol. 4, pp. {Last Accessed March 1, 2022}, 2008.
- [262] C. F. Eyring, "Methods of calculating the average coefficient of sound absorption," *The Journal of the Acoustical Society of America (JASA)*, vol. 4, no. 3, pp. 178–192, 1933.
- [263] C. F. Eyring, "Reverberation time measurements in coupled rooms," *The Journal of the Acoustical Society of America (JASA)*, vol. 3, no. 2A, pp. 181–206, 1931.
- [264] L. L. Beranek, "Analysis of sabine and eyring equations and their application to concert hall audience and chair absorption," *The Journal of the Acoustical Society of America (JASA)*, vol. 120, no. 3, pp. 1399–1410, 2006.
- [265] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, pp. 458–465, 1969.
- [266] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [267] J. D. Markel and A. J. Gray, *Linear Prediction of Speech*. Springer Science & Business Media, 2013, vol. 12.
- [268] A. V. Oppenheim, "Superposition in a class of nonlinear systems," *MIT Research Laboratory of Electronics*, 1965.

- [269] R. W. Schafer, "Echo removal by discrete generalized linear filtering," *MIT Research Laboratory of Electronics*, 1969.
- [270] A. V. Oppenheim, R. W. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 437–466, 1968.
- [271] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun 1968.
- [272] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.
- [273] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco *et al.*, "REDDOTS replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 5-9, 2017, pp. 5395–5399.
- [274] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.
- [275] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.
- [276] F. Hilger and H. Ney, "Quantile-based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 845–854, 2006.
- [277] A. De La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.
- [278] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, United Kingdom, 30-31 August, 2004.
- [279] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.
- [280] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America (JASA)*, vol. 55, no. 6, pp. 1304–1312, 1974.
- [281] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, p. 58, 1996.
- [282] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, 1994.

- [283] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition—general classifier approaches and data fusion methods," *Pattern recognition*, vol. 35, no. 12, pp. 2801–2821, 2002.
- [284] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in asvspoof2017." in *INTERSPEECH*, Stockholm, Sweden, August 20-24, 2017, pp. 87–91.
- [285] H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 726–730.
- [286] N. V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using bayesian framework," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, December 8-12, 2013, pp. 156–161.
- [287] P. G. Radadia and H. A. Patil, "A cepstral mean subtraction based features for singer identification," in *2014 International Conference on Asian Language Processing (IALP)*, Oct. 20-23, 2014, pp. 58–61.
- [288] S. Barathi, S. Yamini, S. Sarkar, and T. Basu, "Comparison of different features for identification of females in multilingual environment," in *Proceedings of the Eleventh National Conference on Communications: NCC-2005, 28-30 January, 2005*, 2005, p. 300.
- [289] "ASVspoof 2019:Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2019. [Online]. Available: http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan/, LastAccess23-Feb-2019
- [290] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.
- [291] E. A. Habets, J. Benesty, S. Gannot, and I. Cohen, "The MVDR beamformer for speech enhancement," in *Speech Processing in Modern Communication*,. Springer, 2010, pp. 225–254.
- [292] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Simon & Schuster, Inc., 1992.
- [293] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.
- [294] M. Karaman, P.-C. Li, and M. O'Donnell, "Synthetic aperture imaging for small scale systems," *IEEE Transactions on Ultrasonic, Ferroelectrics, and Frequency Control*, vol. 42, no. 3, pp. 429–442, 1995.
- [295] L. G. Bezanson, *The Subarray MVDR Beamformer: A Space-Time Adaptive Processor Applied to Active Sonar*. University of California, San Diego, 2013.
- [296] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley & Sons, 2009.
- [297] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

- [298] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.
- [299] J. Wen, “Reverberation: Models, estimation and application,” in *Imperial College London*, 2009.
- [300] 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.
- [301] H. Kotta, A. T. Patil, R. Acharya, and H. A. Patil, “Subband channel selection using TEO for replay spoof detection in voice assistants,” in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 7-10, 2020, pp. 538–542.
- [302] D. B. Freed, *Motor Speech Disorders: Diagnosis and Treatment*, 3.ed. Plural Publishing, USA, 2018.
- [303] P. Swarup, R. Maas, S. Garimella, S. H. Mallidi, and B. Hoffmeister, “Improving ASR confidence scores for alexa using acoustic and hypothesis embeddings,” in *INTERSPEECH*, Graz, Austria, September 15-19, 2019, pp. 2175–2179.
- [304] L. De Russis and F. Corno, “On the impact of dysarthric speech on contemporary ASR cloud platforms,” *Journal of Reliable Intelligent Environments*, vol. 5, no. 3, pp. 163–172, 2019.
- [305] V. Young and A. Mihailidis, “Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review,” *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.
- [306] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, “Severity-based adaptation with limited data for asr to Aid dysarthric speakers,” *Public Library of Science (PLoS) One*, vol. 9, no. 1, 2014.
- [307] C. Bhat, B. Vachhani, and S. K. Kopparapu, “Automatic assessment of dysarthria severity level using audio descriptors,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, March 5-9, 2017, pp. 5070–5074.
- [308] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.
- [309] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [310] X. Zhang and J. Wu, “Deep belief networks based voice activity detection,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.
- [311] X. Zhang and D. Wang, “Boosting contextual information for deep neural network based voice activity detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

- [312] G. Calvert, C. Spence, B. E. Stein *et al.*, *The Handbook of Multisensory Processes*. MIT Press, Edition, 2004.
- [313] D. Sztahó and K. Vicsi, “Estimating the severity of parkinson’s disease using voiced ratio and nonlinear parameters,” in *Statistical Language and Speech Processing*, P. Král and C. Martín-Vide, Eds. Springer International Publishing, 2016, pp. 96–107.
- [314] T. H. Falk, W.-Y. Chan, and F. Shein, “Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility,” *Speech Communication*, vol. 54, no. 5, pp. 622 – 631, 2012.
- [315] M. S. Paja and T. H. Falk, “Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech,” in *INTERSPEECH*, Portland, Oregon, September 9-13, 2012, pp. 62–65.
- [316] N. Gurevich and S. L. Scamihorn, “Speech-language pathologists’ use of intelligibility measures in adults with dysarthria,” *American Journal of Speech-Language Pathology*, vol. 26, no. 3, pp. 873–892, 2017.
- [317] M. M. Hoehn and M. D. Yahr, “Parkinsonism: Onset, Progression, and Mortality,” *Neurology*, vol. 17, no. 5, pp. 427–427, 1967.
- [318] K. Yorkston, D. Beukelman, and C. Traynor, *Computerized Assessment of Intelligibility of Dysarthric Speech*. CC Publications, 1984.
- [319] P. Enderby, “Frenchay dysarthria assessment,” *British Journal of Disorders of Communication*, vol. 15, no. 3, pp. 165–173, 1980.
- [320] S. Fahn and R. Elton, “Unified Parkinson’s Disease Rating Scale (UPDRS),” *Rev Neurol (Paris)*, vol. 156, pp. 534–541, 2000.
- [321] T. Schmitz-Hübsch, S. T. Du Montcel, L. Baliko, J. Berciano, S. Boesch, C. Depondt, P. Giunti, C. Globas, J. Infante, J.-S. Kang *et al.*, “Scale for the Assessment and Rating of Aataxia: Development of a New Clinical Scale,” *Neurology*, vol. 66, no. 11, pp. 1717–1720, 2006.
- [322] K. L. Lansford, V. Berisha, and R. L. Utianski, “Modeling listener perception of speaker similarity in dysarthria,” *The Journal of the Acoustical Society of America (JASA)*, vol. 139, no. 6, pp. EL209–EL215, 2016.
- [323] C. Bhat, B. Das, B. Vachhani, and S. K. Kopparapu, “Dysarthric speech recognition using time-delay neural network based denoising autoencoder.” in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 451–455.
- [324] C. Bhat, B. Vachhani, and S. K. Kopparapu, “Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation.” in *INTERSPEECH*, San Fransisco, USA, Sept. 8-12, 2016, pp. 228–232.
- [325] M. J. Kim, B. Cao, K. An, and J. Wang, “Dysarthric speech recognition using Convolutional LSTM Neural Network.” in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 2948–2952.
- [326] B. Vachhani, C. Bhat, and S. K. Kopparapu, “Data augmentation using healthy speech for dysarthric speech recognition,” in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 471–475.

- [327] H. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 390–399, 2019.
- [328] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *ETRI Journal*, vol. 40, no. 5, pp. 643–652, 2018.
- [329] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth, "A multitask learning approach to assess the dysarthria severity in patients with parkinson's disease," in *INTERSPEECH*, Hyderabad, India, Sept. 2-6, 2018, pp. 456–460.
- [330] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Speech disorder diagnosis and classification," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 1, pp. 57–67, 2014.
- [331] R. de Oliveira Chappaz, S. dos Santos Barreto, and K. Z. Ortiz, "Pneumo-phon-articulatory coordination assessment in dysarthria cases: a cross-sectional study," *São Paulo Med. J.*, vol. 136, no. 3, pp. 216–221, 2018.
- [332] Y. J. Kim, R. D. Kent, and G. Weismer, "An acoustic study of the relationships among neurologic disease, dysarthria type and severity of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 2, pp. 417–429, 2011.
- [333] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 187–194, 2011.
- [334] K. M. Rosen, J. V. Goozee, and B. E. Murdoch, "Examining the effects of multiple sclerosis on speech production: Does phonetic structure matter?" *Journal of Communication Disorders*, vol. 41, no. 1, pp. 49–69, 2008.
- [335] G. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 5, pp. 1001–1013, 1995.
- [336] S. Watanabe, K. Arasaki, H. Nagata, and S. Shouji, "Analysis of dysarthria in amyotrophic lateral sclerosis—MRI of the tongue and formant analysis of vowels," *Rinsho Shinkeigaku*, vol. 34, no. 3, pp. 217–223, 1994.
- [337] O. Christensen and K. L. Christensen, *Approximation Theory: From Taylor Polynomials to Wavelets*. Birkhauser, 2006.
- [338] K. Kawaguchi and Y. Bengio, "Depth with nonlinearity creates no bad local minima in resnets," *Neural Networks*, vol. 118, pp. 167–174, 2019.
- [339] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.
- [340] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

- [341] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015.
- [342] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [343] J. Hestness, N. Ardalani, and G. Diamos, "Beyond human-level accuracy: computational challenges in deep learning," in *the 24th Symposium on Principles and Practice of Parallel Programming*, 2019, pp. 1–14.
- [344] P. Angelov and A. Sperduti, "Challenges in deep learning." in *the 24th European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 2016, pp. 489–496.
- [345] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Honolulu, Hawaii, July 21-26, 2017, pp. 3147–3155.
- [346] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision (ECCV)*, Amsterdam, The Netherlands, October 8-16, 2016, pp. 630–645.
- [347] D. A. Reynolds, "A gaussian mixture modeling approach to text-independent speaker identification." *PhD thesis, Georgia Institute of Technology*, 1992.
- [348] 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.
- [349] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [350] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013, {Last Accessed: Apr 19, 2014}.
- [351] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
- [352] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.
- [353] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *Journal of communication disorders*, vol. 32, no. 3, pp. 141–186, 1999.
- [354] M. Purohit, M. Parmar, M. Patel, H. Malaviya, and H. A. Patil, "Weak speech supervision: A case study of dysarthria severity classification," in *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020, pp. 101–105.
- [355] H. A. Patil, "Cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry," in *A. Newtein (Ed.) Advances in Speech Recognition*, Springer, 2010, pp. 323–348.

- [356] J. J. Engelsma, D. Deb, K. Cao, A. Bhatnagar, P. S. Sudhish, and A. K. Jain, "Infant-id: Fingerprints for global good," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3543–3559, 2021.
- [357] H. A. Patil, "Infant identification from their cry," in *IEEE 7th International Conference on Advances in Pattern Recognition*, Kolkata, India, February 4-6, 2009, pp. 107–110.
- [358] C. C. Onu, I. Udeogu, E. Ndiomu, U. Kengni, D. Precup, G. M. Sant'Anna, E. Alikor, and P. Opara, "Ubenwa: Cry-based diagnosis of birth asphyxia," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, Dec. 4-7, 2017.
- [359] C. C. Onu, J. Lebensold, W. L. Hamilton, and D. Precup, "Neural Transfer Learning for Cry-Based Diagnosis of Perinatal Asphyxia," in *INTERSPEECH*, Graz, Austria, Sept. 15-19, 2019, pp. 3053–3057.
- [360] K. Manickam and H. Li, "Complexity analysis of normal and deaf infant cry acoustic waves," in *Fourth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Firenze, Italy, Oct. 29-31, 2005, pp. 105–108.
- [361] O. Wasz-Höckert, T. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne, "The identification of some specific meanings in infant vocalization," *Experientia*, vol. 20, no. 3, pp. 154–154, 1964.
- [362] Q. Xie, R. K. Ward, and C. A. Laszlo, "Automatic assessment of infants' levels-of-distress from the cry signals," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 253–265, 1996.
- [363] H. F. Alaie, L. Abou-Abbas, and C. Tadj, "Cry-based infant pathology classification using gmms," *Speech Communication*, vol. 77, pp. 28–52, 2016.
- [364] C. Ji, T. B. Mudiyansele, Y. Gao, and Y. Pan, "A review of infant cry analysis and classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–17, 2021.
- [365] G. Gambardella, "Time scaling and short-time spectral analysis," *The Journal of the Acoustical Society of America (JASA)*, vol. 44, no. 6, pp. 1745–1747, 1968.
- [366] A. V. Oppenheim, A. S. Willsky, S. H. Nawab, G. M. Hernández *et al.*, *Signals & Systems*. Pearson Educación, 1997.
- [367] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.
- [368] 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.
- [369] A. Kanervisto, V. Hautamäki, T. Kinnunen, and J. Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 477–488, 2021.
- [370] Z. Rafii, "The constant-q harmonic coefficients: A timbre feature designed for music signals [lecture notes]," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 90–96, 2022.

- [371] P. Gupta, P. K. Chodingala, and H. A. Patil, "Morse wavelet features for pop noise detection," in *accepted in International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, July 11-15, 2022.
- [372] A. Venkitaraman and C. S. Seelamantula, "On computing amplitude, phase, and frequency modulations using a vector interpretation of the analytic signal," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1187–1190, 2013.
- [373] H. Weyl, *The theory of groups and quantum mechanics*. Courier Corporation, 1950.
- [374] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 2, pp. 170–177, 1977.
- [375] J. Tribolet, "Seismic application of homomorphic signal processing," *Ph.D. Dissertation, Massachusetts Institute of Technology (MIT), Cambridge MA, 1977*.
- [376] R. Nobili, F. Mammano, and J. Ashmore, "How well do we understand the cochlea?" *Trends in Neurosciences*, vol. 21, no. 4, pp. 159–167, 1998.
- [377] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. vol. 405, no. 2, pp. pp. 442–451, 1975.

List of Publications from Thesis

Journal Papers

1. Kuldeep Khorja, **Ankur T. Patil**, and Hemant A. Patil, "On Significance of Constant-Q Transform for Pop Noise Detection" in *Computer, Speech & Language*, vol. 77 (2023), pp. 101421.
2. **Ankur T. Patil**, Rajul Acharya, Hemant A. Patil, and Rodrigo Capobianco Guido, "Improving the potential of Enhanced Teager Energy Cepstral Coefficients (ETECC) for replay attack detection" in *Computer, Speech & Language*, Elsevier, vol. 72 (2022), pp. 101281.
3. **Ankur T. Patil**, Hemant A. Patil, and Kuldeep Khorja, "Effectiveness of Energy Separation-Based Instantaneous Frequency Estimation for Cochlear Cepstral Features for Synthetic and Voice-converted Spoofed Speech Detection" in *Computer, Speech & Language*, Elsevier, vol. 72 (2022), pp. 101301.
4. Siddhant Gupta, **Ankur T. Patil**, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A. Patil, and Rodrigo C. Guido, "Residual Neural Network Precisely Quantifies Dysarthria Severity-level based on Short-duration Speech Segments", in *Neural Networks*, Elsevier, 139(2021): 105-117.
5. **Ankur T. Patil**, Anand Therattil, and Hemant A. Patil, "On Significance of Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants," revised and resubmitted in *Computer, Speech & Language*, Elsevier.

Book Chapters

1. Aastha Kachhi, Anand Therattil, **Ankur T. Patil**, Hardik B. Sailor, Hemant A. Patil, "Significance of Energy Features for Severity-Level Classification of Dysarthria" in: *International Conference on Speech and Computer (SPECOM), Lecture Notes in Computer Science (LNAI)*, vol 13721, pp. 325-337, November 2022, Springer, Cham.
2. **Ankur T. Patil**, Harsh Kotta, Rajul Acharya, and Hemant A. Patil, "Spectral Root Features for Replay Spoof Detection in Voice Assistants," in: *International Conference on Speech and Computer (SPECOM), Lecture Notes*

in Computer Science (LNAI), vol 12997, pp. 504-515, Sept. 2021, Springer, Cham.

Conference Papers

1. Aastha Kachhi, Anand Therattil, **Ankur T. Patil**, Hardik B. Sailor, Hemant A. Patil, "Teager Energy Cepstral Coefficients For Classification of Dysarthric Speech Severity-Level" in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 2022, pp. 1462-1468.
2. Piyushkumar Chodingala, Shreya Chaturvedi, **Ankur T. Patil**, and Hemant A. Patil, "Robustness of DAS Beamformer Over MVDR for Replay Attack Detection On Voice Assistants," in IEEE International Conference on Signal Processing and Communications (SPCOM)-2022, Bangalore, India, July 11-15, 2022, pp. 1-5.
3. **Ankur T. Patil**, Kuldeep Khorria, Hemant A. Patil, "Voice Liveness Detection using Constant-Q Transform-Based Features," in European Signal Processing Conference (EUSIPCO)-2022, Belgrade, Serbia, August 29 - Sept. 2, 2022, pp. 110-114.
4. **Ankur T. Patil**, Aastha Kachhi, Hemant A. Patil, "Subband Teager Energy Representations for Infant Cry Analysis and Classification," in European Signal Processing Conference (EUSIPCO)-2022 Belgrade, Serbia, August 29 - Sept. 2, 2022, pp. 1313-1317.
5. Hemant A. Patil, Rajul Acharya, **Ankur T. Patil**, Priyanka Gupta, "Non-Cepstral Uncertainty Vector for Replay Spoofed Speech Detection," in European Signal Processing Conference (EUSIPCO)-2022 Belgrade, Serbia, August 29 - September 2, 2022, pp. 374-378.
6. Hemant A. Patil, **Ankur T. Patil**, Aastha Kachhi, "Constant Q Cepstral Coefficients for Classification of Normal *vs.* Pathological Cry," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, May 7-13, 2022, pp 7392-7396.
7. Siddhant Gupta, Kuldeep Khorria, **Ankur T. Patil**, and Hemant A. Patil, "Deep Convolutional Neural Network for Voice Liveness Detection," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Tokyo, Japan, December 14-17, 2021, pp. 775-779.

8. Rajul Acharya, Harsh Kotta, **Ankur T. Patil**, and Hemant A. Patil, "Cross-Teager Energy Cepstral Coefficients for Replay Spoof Detection on Voice Assistants," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Ontario, Canada, June 6-11, 2021, pp. 6364-6368.
9. Kuldeep Khorra, **Ankur T. Patil**, and Hemant A. Patil, "Significance of Constant-Q Transform for VoiceLiveness Detection," in European Signal Processing Conference (EUSIPCO), Dublin, Ireland, August 2021, pp. 126-130.
10. **Ankur T. Patil**, and Hemant A. Patil, "Significance of CMVN for Replay Spoof Detection," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Auckland, New Zealand, Dec. 7-10, 2020, pp. 532-537.
11. Harsh Kotta, **Ankur T. Patil**, Rajul Acharya, and Hemant A. Patil, "Subband Channel Selection using TEO for Replay Spoof Detection in Voice Assistants," In 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, December 7-10, 2020, pp. 538-542.
12. Mirali Purohit, Maitreya Patel, Harshit Malaviya, **Ankur T. Patil**, Mihir Parmar, Nirmesh Shah, Savan Doshi, and Hemant A. Patil, "Intelligibility Improvement of Dysarthric Speech using MMSE DiscoGAN," In 2020 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, July 19-24, 2020, pp. 1-5.
13. Madhu R. Kamble, Pulikonda Aditya Krishna Sai, Maddala V. Siva Krishna, **Ankur T. 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, November 18-21, 2019, pp. 1545-1550.
14. **Ankur T. Patil**, Rajul Acharya, Pulikonda Krishna Aditya Sai, and Hemant A. Patil, "Energy Separation-Based Instantaneous Frequency Estimation for Cochlear Cepstral Feature for Replay Spoof Detection," in INTERSPEECH, Graz, Austria, September 15-19, 2019, pp. 2898-2902.
15. Prasad A. Tapkir, **Ankur T. Patil**, Neil Shah, and Hemant A. Patil, "Novel spectral root cepstral features for replay spoof detection," in 2018 Asia-Pacific

Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, Hawaii, USA, November 12-15, 2018, pp. 1945-1950.

16. **Ankur T. Patil**, Maddala V. Siva Krishna, Mehak Piplani, Pulikonda Aditya Sai, Hardik B. Sailor, and Hemant A. Patil. "DA-IICT/IITV System for the 5th CHiME 2018 Challenge," The 5th International Workshop on Speech Processing in Everyday Environments (CHiME-2018), Hyderabad, India, September 7, 2018.
17. Hardik B. Sailor, Maddala Venkata Siva Krishna, Diksha Chhabra, **Ankur T. Patil**, Madhu R. Kamble, and Hemant A. Patil, "DA-IICT/IITV System for Low Resource Speech Recognition Challenge 2018," in INTERSPEECH, Hyderabad, India, September 2-6, 2018, pp. 3187-3191.
18. Hardik B. Sailor, **Ankur T. Patil**, and Hemant A. Patil, "Advances in Low Resource ASR: A Deep Learning Perspective," In Spoken Language Technologies for Under-resourced Languages (SLTU), Delhi, India, August 29-31, 2018, pp. 15-19.
19. **Ankur T. Patil**, Hemant A. Patil, "On applicability of the CMVN for genuine vs. replay spoof speech detection" rejected in APSIPA-ASC-2022.

Brief Biography



Ankur T. Patil received B.E. degree from D. N. Patel CoE, Shahada, Nandurbar, Maharashtra, India in 2009. He did M.E. in 2015 from K. K. Wagh IEER, Nashik, Maharashtra, India. He was a doctoral student during July-2016 to July-2022 under supervision of Prof. (Dr.) Hemant A. Patil at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India. Currently, he is working as an Audio Data Scientist at DADJ Inc.

His main research is focused on developing signal processing-based countermeasures and analysis of natural *vs.* spoof speech signals. Furthermore, he also contributed in various speech technology applications, namely, severity-level classification of dysarthric speech and its enhancement, classification of normal *vs.* pathological infant cries and automatic speech recognition (ASR). He was a research intern at Samsung Research Institute, Bangalore (SRI-B), India during May-July, 2019. He received student travel grant of 150 Euros for presenting his paper in the 5th CHiME Speech Separation and Recognition Challenge, 2018 in Hyderabad, India. He also attended INTERSPEECH-2018 and 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)-2018. He is a student member of IEEE, and IEEE Signal Processing Society. He has been a student participant and volunteer for ISCA supported Summer Schools such as, S4P 2019, S4P 2018, and S4P 2017.

In DA-IICT, he was teaching assistant for various courses such as, Deep Learning, Analog & Digital Communication, Signals & Systems, Digital Signal Processing, Advanced Digital Signal Processing, Cyber Physical Systems and IoT, etc. Before joining to DA-IICT as a research scholar, he was lecturer in MET BKC IOE, Nashik, Maharashtra and has taught various courses such as, Signals & Systems,

Electromagnetics, Basic Electronics, Power Electronics, Wave theory & Antenna,
etc.