

Features for Live and Spoofed Speech Detection

by

Priyanka Gupta
201721001

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




December 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.

 (27/12/2023)

Ms. Priyanka Gupta
(Student ID : 201721001)

Certificate

This is to certify that the thesis work entitled, "*Features for Live and Spoofed Speech Detection*," has been carried out by Ms. Priyanka Gupta for the degree of Doctor of Philosophy at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) under my supervision.


27/12/2023

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

Acknowledgments

There are no proper words to convey my deep gratitude and respect for my thesis supervisor, Prof. (Dr.) Hemant A. Patil for his dedicated help, encouragement, and continuous support throughout my Ph.D. His dedication to publish high-quality research work in top conferences and peer reviewed journals has made a deep impression on me. His dedication towards his students is unmatched, wherein he was fighting for his life in the hospital due to the pandemic, and was still concerned about his students' work in the lab. It was due to his support that I could overcome even non-academic hurdles during my Ph.D., such as the pandemic and my health issues. It was his belief in me that kept my determination alive even in the hardest of days! 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.) M. V. Joshi and Prof. (Dr.) Rajib Lochan Das for their valuable and eye-opening insights and feedback, which helped me plan the further course of my work, at the end of each semester.

I am grateful to the organizers of EUSIPCO 2023 for giving me the opportunity to be the Session Chair in EUSIPCO 2023. I also thank them for the EURASIP Student Travel Grant of 700 Euros. I also gratefully acknowledge the partial support from the Ministry of Electronics and Information Technology (MeitY), Govt. of India 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 cooperation.

My heartfelt thanks to all my co-authors, Piyush K. Chodingala, Anand Therattil, Aastha Kacchi, Rajul Acharya, Dr. Ankur T. Patil, Siddhant Gupta, Shrishti Singh, Gauri P. Prajapati, Dr. Madhu R. Kamble, for their help, support, and technical discussions during this research work. I would also like to thank other members of the Speech Research Lab @ DA-IICT Mirali Purohit, Divyesh Rajpura, Kuldeep Khorla, Dipesh Singh, and Shreya Chaturvedi, and all my co-authors for

having a friendly and conducive environment in the lab.

This work would not have been possible without the support of my friends, who were there to listen to me during my emotional turmoils. Special thanks to Mr. Sameer More, who played a small yet significant role by listening to my non-stop rants during my low times.

I am more than grateful to my dear parents (Mr. Yogendra Kumar Gupta, and Mrs. Sanju Gupta) who were there with me in every hurdle-big or small, in every moment of despair as well as joy. Their patience and constant motivations have always been my strength and will be the greatest lessons of my life, forever.

Priyanka Gupta

Contents

Abstract	xii
List of Acronyms	xiii
List of Symbols	xv
List of Figures	xvi
List of Tables	xxiv
1 Introduction	1
1.1 Motivation	2
1.2 Research Objectives	4
1.2.1 Replay Spoofed Speech Detection (SSD) Problem	4
1.2.2 Voice Liveness Detection (VLD) Problem	4
1.3 Contributions and Scope of the Thesis	5
1.3.1 Features for Replay SSD	6
1.3.2 Features for Voice Liveness Detection (VLD)	7
1.3.3 Voice Privacy and Attacker’s Perspective	7
1.3.4 Additional Applications	8
1.4 Organization of the Thesis	8
1.5 Chapter Summary	9
2 Literature Survey	11
2.1 Introduction	11
2.2 Replay Spoofed Speech Detection (SSD)	11
2.3 Voice Liveness Detection (VLD)	15
2.4 Studies from the Attacker’s Perspective	17
2.5 Research Gaps and Contributions of the Thesis	24
2.6 Chapter Summary	27
3 Experimental Setup	29
3.1 Introduction	29
3.2 Standard Corpora Used For Anti-Spoofing	29

3.2.1	ASVSpooof 2015 Challenge Dataset	30
3.2.2	ASVSpooof 2017 Challenge Dataset	31
3.2.3	ASVSpooof 2019 Challenge Dataset	32
3.2.3.1	Logical Access (LA)	33
3.2.3.2	Physical Access (PA)	34
3.2.4	ASVSpooof 2021 Challenge Dataset	34
3.2.5	Biometrics: Theory, Applications, and Systems (BTAS) 2016 Dataset	35
3.2.6	Realistic Replay Attack Microphone Array Speech Corpus (ReMASC)	35
3.3	Standard Corpora used For Voice Liveness Detection (VLD)	37
3.3.1	POp noise COrpus (POCO)	37
3.4	Classifiers Used	39
3.4.1	Gaussian Mixture Model (GMM)	39
3.4.2	Convolutional Neural Network (CNN)	40
3.4.3	Light Convolutional Neural Network (LCNN)	41
3.4.4	Residual Neural Network (ResNet)	41
3.5	Performance Metrics Used	42
3.6	Score-Level Data Fusion	43
3.7	Chapter Summary	43
4	Features for Replay Spoofed Speech Detection	45
4.1	Introduction	45
4.2	CFCCIF-QESA	47
4.2.1	Motivation for CFCCIF-QESA	47
4.2.2	Estimation of Instantaneous Frequency (IF)	47
4.2.2.1	IF Estimation Using Analytic Signal	47
4.2.2.2	IF Estimation Using ESA	48
4.2.3	Proposed CFCCIF-QESA Feature Set	49
4.2.3.1	Optimal Relative Phase using MI	49
4.2.3.2	Incorporation of Quadrature-Phase Component	53
4.2.3.3	TEO for Complex-Valued Signal	54
4.2.3.4	Proposed Quadrature ESA (QESA)	55
4.2.3.5	Alleviation of Some of the difficulties Associated With IF	56
4.2.3.6	CFCCIF-QESA Feature Extraction	57
4.2.3.7	Spectrographic Analysis of CFCCIF-ESA <i>vs.</i> CFCCIF- QESA	61

4.2.4	Setup	61
4.2.5	Experimental Results	63
4.2.5.1	Initial Parameterization	63
4.2.5.2	Parameter Tuning on the ASVSpooF 2017 v2.0 Dataset	63
4.2.5.3	Results on the ASVSpooF 2017 v2.0 Database w.r.t. Various Classifiers	66
4.2.5.4	Analysis of Latency on the ASVSpooF 2017 v2.0 Database	68
4.2.5.5	Cross-Database Evaluation with Training on ASVSpooF 17 V2.0	70
4.2.5.6	SSD System Performance Under Ideal Conditions .	72
4.2.5.7	Results on the ASVSpooF 2019 PA Database	73
4.2.5.8	Results on the BTAS 2016 Dataset	73
4.2.5.9	Results on the ReMASC Dataset	74
4.2.5.10	Results on the ASVSpooF 2015 dataset	76
4.2.5.11	Analysis Using Model-Level Measures	79
4.3	Optimized Linear Frequency Residual Cepstral Coefficients (LFRCC)	84
4.3.1	Linear Prediction (LP)	85
4.3.2	Proposed Optimized LFRCC	86
4.3.3	Setup	88
4.3.4	Experimental Results	89
4.4	U-Vector	90
4.4.1	Time-Bandwidth Product (TBP)	91
4.4.2	U-Vector Feature Extraction	93
4.4.3	Setup	96
4.4.4	Experimental Results	96
4.5	Chapter Summary	98
5	Features for Voice Liveness Detection (VLD)	101
5.1	Introduction	101
5.2	CWT-Based Approach	102
5.3	Bump Wavelet-Based Features	105
5.3.1	Proposed Approach	105
5.3.2	Setup	106
5.3.3	Speaker-Microphone Distance-Based Analysis	108
5.3.4	Experimental Results	111
5.4	Morlet Wavelet-Based Features	115
5.4.1	Proposed Approach	115

5.4.1.1	Handcrafted Morlet Wavelet-Based Features	115
5.4.1.2	Low Frequency Morlet Scalogram-Based Features	117
5.4.2	Setup	117
5.4.3	Speaker-Microphone Distance-Based Analysis	117
5.4.4	Experimental Results	120
5.4.4.1	Proposed Handcrafted Morlet-Based Features . . .	120
5.4.4.2	Proposed Morlet Scalogram-Based Features	121
5.5	Generalized Morse Wavelet (GMW)-Based Features	125
5.5.1	Proposed Approach	125
5.5.1.1	Generalized Morse Wavelets (GMWs)	125
5.5.1.2	Advantage of GMW	129
5.5.1.3	Morse Wavelet-Based Features for the VLD task . .	130
5.5.2	Setup	131
5.5.3	Speaker-Microphone Distance-Based Analysis	134
5.5.4	Experimental Results	136
5.5.4.1	Effect of $P_{\beta,\gamma}^2 = \beta\gamma$	136
5.5.4.2	Effect of γ	137
5.5.4.3	Effect of Frequency Range	139
5.5.4.4	Phoneme-Based Analysis	139
5.5.4.5	Effect of Distance Between Genuine Speaker and Attacker's Microphone	144
5.5.4.6	Effect of Classifier Structure	145
5.5.4.7	Performance Under Ideal Conditions	146
5.5.4.8	Performance Comparison With End-To-End Neu- ral Network Model	146
5.6	Chapter Summary	147
6	Voice Privacy and Attacker's Perspective	149
6.1	Introduction	149
6.1.1	Motivation for Voice Privacy	150
6.1.2	De-identification <i>vs.</i> Anonymization	151
6.2	Voice Privacy Using Linear Prediction (LP) Model	152
6.2.1	Speech Production Model w.r.t. Linear Acoustics	152
6.2.2	Energy Losses	156
6.2.3	Linear Prediction (LP) Model	159
6.2.4	Proposed Voice Privacy System	162
6.3	Experimental Setup	163
6.3.0.1	Datasets Used	163

6.3.1	Experimental Results	163
6.3.1.1	Gender-Based Analysis	164
6.4	Voice Privacy and Attacker’s Perspective	167
6.4.1	Target Selection	167
6.4.2	Target Selection by the Attacker and Voice Privacy System .	169
6.5	Target Selection in Enrolled Users with Malicious Intent	171
6.5.1	Setup	172
6.5.2	Experimental Results	173
6.6	Technological Challenged Faced By the Attacker	174
6.6.1	Number of Trials on Victim ASV Access	174
6.6.2	Corpora for Attacker’s Perspective	174
6.6.3	Transmission Channel	175
6.6.4	Perturbation Minuteness in Adversarial Attacks	176
6.6.5	Voice Privacy Systems	176
6.6.6	Voice Liveness Detection (VLD)	177
6.6.7	DeepFake Detectors	177
6.7	Voice Privacy and Cryptography	178
6.7.1	Public Key Encryption	178
6.7.2	Limitations of Cryptographic Approaches for Voice Privacy	180
6.8	Chapter Summary	181
7	Additional Works	183
7.1	Infant Cry Classification	183
7.1.1	Morse wavelets-Based Features	185
7.1.1.1	Experimental Setup	187
7.1.1.2	Experimental Results	187
7.1.2	Uncertainty Feature Vector	190
7.1.2.1	Experimental Setup	192
7.1.2.2	Experimental Results	192
7.2	Dysarthric Severity-Level Classification	195
7.2.1	Morse wavelet-based features	197
7.2.1.1	Experimental Setup	198
7.2.1.2	Experimental Results	199
7.3	Chapter Summary	200
8	Summary and Conclusions	201
8.1	Summary of the Thesis	201
8.2	Limitations of This Work	204
8.3	Future Research Directions	204

8.4 Open Research Problems 205

Appendix A Analytic Signal 207

Appendix B Teager Energy Operator (TEO) 209

Appendix C Modelling Speech as an AM-FM Signal 211

Appendix D IF Estimation using ESA 213

**Appendix E Heisenberg’s Uncertainty Principle in Signal Processing Frame-
work 215**

Appendix F Energy Conservation of Time-Frequency Transforms 217

List of Publications from the Thesis 240

Abstract

The authorization to access specific information is given by a biometric system. Biometric systems are used for security purposes in a way that they prevent unauthorized access to important information or data (*information privacy*). The access granted by the biometric is done by capturing traits of humans, which make all human beings unique w.r.t. that particular trait. This thesis focuses on voice-based biometric systems, also known as Automatic Speaker Verification (ASV) systems, given that speech is the most natural and powerful form of communication used by humans to communicate with the outside world. It is the most intuitive, simple, and easy-to-produce characteristic. Since ASV systems have been used for applications, such as in banking transactions and access to buildings associated with classified information, only authorized legitimate or *genuine* users are granted access.

ASV systems suffer from vulnerabilities to attacks and can be compromised at various stages. The attacks may be categorized as direct and indirect attacks, depending on the extent of the attacker's accessibility to the ASV framework. Besides, due to the recent commercial success of several Intelligent Personal Assistants (IPAs), also known as voice assistants, such as Speech Interpretation and Recognition Interface (SIRI), Amazon Alexa, Google Home, and so on, many voice-enabled devices in Internet of Things (IoT) have been commonly prone to spoofing attacks. To that effect, there is active research in the direction of designing countermeasure systems for ASV systems, particularly for spoofing attacks, namely, Speech Synthesis (SS), Voice Conversion (VC), and replay.

This thesis is a humble attempt to alleviate some of the research gaps in designing features for countermeasure systems. In particular, this thesis proposes Quadrature Energy Separation Algorithm (QESA) in the light of incorporating the quadrature-phase component with the in-phase component of the signal. To that effect, an existing feature set for replay Spoofed Speech Detection (SSD), namely, CFCCIF-ESA is extended to the CFCCIF-QESA feature set for enhanced performance of the countermeasure system. The performance of the proposed CFCCIF-QESA feature set is evaluated on various datasets for various spoofing attacks

given in the literature. Furthermore, the existing Linear Frequency Residual Cepstral Coefficients (LFRCC) feature set is optimized w.r.t. to its Linear Prediction (LP) order for the replay SSD task. In particular, it is found that the LP order needed for a good prediction of speech is not the same as that needed for the replay SSD task. The resulting *optimized* LFRCC feature set is evaluated on the ASVSpooof 2019 PA dataset. In addition to this, another feature, known as the uncertainty vector (u-vector), is developed from the Heisenberg's uncertainty principle in the signal processing framework. The proposed u-vector is evaluated using the ASVSpooof 2017 dataset for replay attacks.

Furthermore, in the direction to make countermeasure systems *independent* of the type of spoofing attack, features have been proposed for the Voice Liveness Detection (VLD) task. VLD is performed by the detection of pop noise which is the discriminating acoustic cue present in live speech, produced due to the breathing effect captured by the microphone when the speaker's mouth is close to the microphone. The work on VLD in this thesis is based on two key hypotheses, namely, Parseval's energy equivalence for STFT, CWT, and analytic CWT, whereas the second hypothesis is that the energy of pop noise decreases with the distance of a microphone from the speaker that is used to capture genuine speech. The proposed features for VLD in this thesis are wavelet-based, wherein three wavelets are used, namely, Bump, Morlet, and Morse wavelet, where Morse wavelet is presented as a superfamily of analytic wavelets, called as Generalized Morse Wavelets (GMWs). Detailed experimental analysis such as speaker-microphone proximity, the effect of phoneme type, and the effect of frequency range is studied.

Apart from this, the security of speech data is also taken into account and this thesis proposes an improved Voice Privacy (VP) system, which is based on Linear Prediction (LP) of speech. Furthermore, the VP system is studied along with the attacker's perspective using the target selection approach, and particularly, target selection w.r.t. twins is studied, wherein the most vulnerable twin-pair (i.e., target) is selected. Lastly, some of the proposed feature sets in this thesis are also evaluated for tasks related to other Assistive Speech Technologies (AST) applications, such as the classification of healthy *vs.* pathological infant cries, and dysarthric severity-level classification.

List of Acronyms

AST	Assistive Speech Technologies
ASV	Automatic Speaker Verification
CFCCIF	Cochlear Filter Cepstral Coefficients Instantaneous Frequency
CM	Countermeasures
CMC	Constant-Q Multi-level Coefficients
CQCC	Constant-Q Cepstral Coefficients
CQT	Constant-Q Transform
CWT	Continuous Wavelet Transform
DF	DeepFake
FGSM	Fast Gradient Sign Method
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
GMW	Generalized Morse Wavelet
HE	Homomorphic Encryption
HPF	Highpass Filter
IoT	Internet of Things
IPA	Intelligent Personal Assistant
IPA	International Phonetic Alphabet
LA	Logical Access
LCNN	Light Convolutional Neural Network

LDS	Local Distribution Smoothness
LFRCC	Linear Frequency Residual Cepstral Coefficients
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LPMS	Log Power Magnitude Spectrum
LSTM	Long Short-Term Memory
MGDCC	Modified Group Delay Cepstral Coefficients
MLT	Multi-Level Transform
PA	Physical Access
PPG	Phonetic Posteriorgram
QESA	Quadrature Energy Separation Algorithm
RIR	Room Impulse Response
SIRI	Speech Interpretation and Recognition Interface
SS	Speech Synthesis
SSD	Spoofed Speech Detection
STFT	Short-Time Fourier Transform
t-DCF	tandem Detection Cost Function
TTS	Text-To-Speech
U-Vector	Uncertainty Vector
VAE	Variational Autoencoder
VC	Voice Conversion
VCTK	Voice Cloning ToolKit
VLD	Voice Liveness Detection

List of Symbols

$p(\cdot)$	Probability Density Function
t	Time
ω	Frequency
$x(t)$	Continuous-Time Speech Signal
$x(n)$	Discrete-Time Speech Signal
$H(z)$	System Function in Z-domain
$H(e^{j\omega})$	Frequency Response of System
$\psi\{\cdot\}$	Teager Energy Operator
$\psi_c\{\cdot\}$	Teager Energy Operator for Complex-Valued Signals
$a_i(n)$	Instantaneous Amplitude
ω_i	Instantaneous Frequency
$\phi(t)$	Instantaneous Phase
*	Convolution Operation
Δ	Delta or Dynamic or Velocity Features
$\Delta\Delta$	Delta-Delta or Double-Delta or Acceleration Features
$\psi(t)$	Wavelet in Time-domain
$\Psi(\omega)$	Wavelet in Frequency-domain
α	Morse Wavelet Family Parameter
γ	Morse Wavelet Skewness Parameter
$P_{\beta,\gamma}^2$	Morse Wavelet Duration Parameter
C^∞	Space of Infinitely Differentiable Functions
$L^2(\mathcal{R})$	Hilbert Space of Square Integrable Functions
F_0	Fundamental Frequency
μ	Mean
σ	Variance
u	Uncertainty
λ_n	GMM for Natural Speech
$\rho(\cdot)$	Softmax Function

List of Figures

1.1	A conventional voice biometric (ASV) system. After [1,2].	2
1.2	Spoofed Speech Detection (SSD) system for ASV system.	4
1.3	Safeguarding an ASV system (a) using a replay SSD system, (b) using a VLD system. Best viewed in color.	5
1.4	Organization of the thesis.	8
2.1	A selected chronological progress depicting the development from auditory transform to the cochlear filter-based feature sets.	13
2.2	A selected chronological progress depicting the development of VLD systems, and the applications of Morlet wavelet in the literature.	16
2.3	Classifying various attacks on an ASV system.	17
3.1	The microphone array consists of 15 Audio-Technica AT9903 microphones (M1 to M15) without pop filter. Speaker’s mouth is positioned in front of mic M7 at a distance of d cm from the mic M7.	37
3.2	Conceptual functional block diagram of CNN. After [3].	41
3.3	Conceptual functional block diagram of ResNet. After [3].	42
4.1	Flowchart of the contents of this Chapter w.r.t. the proposed features for the replay SSD task.	46
4.2	(a) AM-FM signal, and (b) MI between AM-FM signal and its phase-shifted version.	51
4.3	(a) Cosine signal, and (b) MI between cosine signal and its phase-shifted version.	51
4.4	Functional block diagram for incorporation of quadrature-phase component. After [4].	53
4.5	Functional block diagram of the proposed CFCCIF-QESA feature set, along with the conventional CFCCIF and CFCCIF-ESA feature sets. The analytic signal in the dotted box is generated w.r.t. procedure shown in Figure 4.4.	57

4.6	Spectrographic representation of the genuine <i>vs.</i> spoofed speech. Panel I and Panel II represent the spectrographic representation of CFCCIF-ESA and CFCCIF-QESA, respectively. Here, (a) genuine speech signal, and (b) corresponding spoofed (replay) speech signal.	60
4.7	Panel I and Panel II represent the analysis of waterfall plots for CFCCIF-ESA and CFCCIF-QESA, respectively, for the same utterances used in Figure 4.6. Here, (a) and (b) represent the waterfall plots for genuine speech signal, and (c) and (d) represent the waterfall plots for spoofed speech signal.	60
4.8	Results (in % EER) for the CFCCIF-QESA feature set on Dev and Eval set of ASVSpooF 2017 v2.0 dataset w.r.t. the (a) number of subband filters in the cochlear filterbank, (b) dimension of feature vector, (c) value of α , (d) value of β , and (e) number of mixtures in GMM.	64
4.9	Analysis of latency period for the SSD system (a) Dev set, and (b) Eval set of ASVspooF 2017 dataset using various feature sets.	69
4.10	Results (in % EER) on Dev set of BTAS 2016 dataset with variation of (a) value of α , and (b) value of β	74
4.11	DET curves on (a) Dev set, and (b) Eval set of ReMASC dataset.	76
4.12	Latency curves for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA on ReMASC dataset.	76
4.13	DET curves (a) Dev, and (b) Eval set.	78
4.14	Latency curves (a) Dev, and (b) Eval set of the ASVSpooF 2015 dataset.	79
4.15	Comparative analysis of KLD and JSD for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA for various numbers of mixtures used in GMM on the ASVSpooF 2017 training corpus.	80
4.16	Comparative analysis of KLD and JSD for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA for various numbers of mixtures used in GMM on the ASVSpooF 2019 PA training corpus.	80
4.17	Comparative analysis of KLD and JSD for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA for various numbers of mixtures used in GMM on BTAS 2016 training corpus.	81
4.18	Comparative analysis of KLD and JSD for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA for various numbers of mixtures used in GMM on VSDC training corpus (only 0PR-1PR).	82

4.19	Comparative analysis of KLD and JSD for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA for various numbers of mixtures used in GMM on VSDC training corpus (only 0PR-2PR).	83
4.20	Comparative analysis of KLD and JSD for CFCCIF-ESA <i>vs.</i> CFCCIF-QESA for various numbers of mixtures used in GMM on ReMASC training corpus.	84
4.21	Plots of framewise magnitude spectrum $ R(\omega) $ of the LP residual of genuine speech, for different values of LP order p	87
4.22	Functional Block diagram of LFRCC Feature Extraction. After [5].	87
4.23	Functional block diagram of the proposed t -vector, ω -vector, and u -vector.	93
4.24	Representation of a genuine speech signal (Panel-I) <i>vs.</i> representation of a spoof speech signal (Panel-II): (a) speech signal, (b) t-gram, (c) ω -gram, (d) u-gram, (e) CQT-gram, (f) spectrogram, and (g) Mel-spectrogram. Ordinate values of the Panel-II subplots are similar to that of Panel-I. Abscissa values of 4.24 (a)-(f) are similar to that of Figure 4.24 (g).	94
4.25	Variation of % EER <i>vs.</i> window size (ms) for u -vector.	96
4.26	DET curves for replay SSD system. The individual DET curves for u -vector (proposed), t -vector, ω -vector, CQCC, MFCC, LFCC on (a) Dev set, and (b) Eval set.	98
5.1	Tiling of the Time-frequency plane for STFT <i>vs.</i> CWT. After [6].	103
5.2	Flowchart of the contents of this Chapter w.r.t. the Proposed CWT-Based Features for VLD.	104
5.3	Panel I represents the case of presence of pop noise (genuine or live speech). Panel II represents suppressed pop noise (spoofed speech) due to the use of pop filter. (a) Time-domain signal for the word 'thong', (b) corresponding scalogram, and (c) selected region of scalogram in (b) corresponding low-frequency (0 – 40 Hz). Solid boxes in Panel I indicates the presence of pop noise, while corresponding dotted boxes in Panel II indicates that the pop noise has been suppressed by the use of pop filter.	106
5.4	Proposed Approach for the VLD task.	106
5.5	The CNN architecture used for classification of the proposed bump wavelet-based scalogram features. After [7].	108

5.6	Panel I, Panel II, and Panel-III represent the varying distance of the speaker from the microphone, i.e., 5 cm, 5.39 cm, and 6.42 cm, respectively, for (a) time-domain signal for the word ‘ <i>dad</i> ’, (b) corresponding scalogram, and (c) selected region of scalogram in (b) corresponding to low-frequency (0 – 40 Hz). Solid boxes in red indicate the presence of pop noise.	109
5.7	Pop noise energies of various phonemes plotted w.r.t. the distance of the speaker from various microphones for the case, when the speaker is at a distance of 5 cm from Mic 7. The pop noise energies are obtained using the proposed Bump wavelet-based approach given via Algorithm 5. The trendlines in each of the sub-figures indicate that the energy of pop noise decreases with the distance of the speaker’s mouth from the microphone.	111
5.8	Word-wise VLD accuracy of STFT-based baseline method <i>vs.</i> proposed bump wavelet-based feature.	112
5.9	Phoneme-wise analysis of genuine (live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Bump wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is present, for (a) plosive (the sample word is ‘ <i>tip</i> ’), (b) whisper (the sample word is ‘ <i>who</i> ’), (c) fricative (the sample word is ‘ <i>laugh</i> ’), (d) affricate (the sample word is ‘ <i>chip</i> ’), (e) nasal (the sample word is ‘ <i>arm</i> ’), and (f) liquid (the sample word is ‘ <i>run</i> ’). . .	113
5.10	Phoneme-wise analysis of spoofed (non-live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Bump wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is absent, for (a) plosive (the sample word is ‘ <i>tip</i> ’), (b) whisper (the sample word is ‘ <i>who</i> ’), (c) fricative (the sample word is ‘ <i>laugh</i> ’), (d) affricate (the sample word is ‘ <i>chip</i> ’), (e) nasal (the sample word is ‘ <i>arm</i> ’), and (f) liquid (the sample word is ‘ <i>run</i> ’).	114
5.11	Panel I represent the case of presence of pop noise (genuine or live speech). Panel II represents suppressed pop noise (spoofed speech) due to the use of pop filter: (a) time-domain signal for the word ‘ <i>laugh</i> ’, (b) corresponding Morlet wavelet-based scalogram, and (c) selected region of scalogram in (b) corresponding low-frequency (0 – 40 Hz). Solid boxes in Panel I indicate the presence of pop noise, while corresponding dotted boxes in Panel II indicates that the pop noise has been suppressed by the use of pop filter.	116

5.12	Panel I, Panel II, and Panel-III represent the varying distance of the speaker from the microphone, namely, 5 cm, 10.78 cm, and 20.40 cm, respectively, for (a) time-domain signal for the word ‘pink’, (b) corresponding scalogram, and (c) selected region of scalogram corresponding low-frequency (0 – 40 Hz). Solid boxes in red indicate the presence of pop noise.	118
5.13	Pop noise energies for various phoneme sounds plotted w.r.t. the distance of the speaker from various microphones for the case when the speaker is at a distance of 5 cm from Mic 7. The pop noise energies are obtained using the proposed Morlet wavelet-based Algorithm 7. The dotted curve in each of the sub-figures indicates that the energy of pop noise decreases with the distance of the speaker’s mouth from the microphone.	119
5.14	Phoneme-wise analysis of genuine (live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morlet wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is present, for (a) plosive (the sample word is ‘tip’), (b) fricative (the sample word is ‘laugh’), (c) whisper (the sample word is ‘who’), (d) nasal (the sample word is ‘arm’), (e) liquid (the sample word is ‘run’), and (f) affricate (the sample word is ‘chip’).	122
5.15	Phoneme-wise analysis of spoofed (non-live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morlet wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is absent, for (a) plosive (the sample word is ‘tip’), (b) fricative (the sample word is ‘laugh’), (c) whisper (the sample word is ‘who’), (d) nasal (the sample word is ‘arm’), (e) liquid (the sample word is ‘run’), and (f) affricate (the sample word is ‘chip’).	123
5.16	Word-wise VLD accuracies (in %) with CNN classifier for (C): Full frequency spectrogram, (D): Low frequency Mel-spectrogram, (E): Handcrafted Bump wavelet-based features, (F): Handcrafted Morlet wavelet-based features, and (G): Handcrafted Morlet scalogram. The indices (C)-(G) are w.r.t. the labels in Table 5.3.	124
5.17	Morse wavelets for varying values of β and γ in (a) time-domain, and (b) frequency-domain. After [8].	126
5.18	Effect of γ parameter on the time-frequency Heisenberg area $A_{\beta,\gamma}$ w.r.t. wavelet duration $P_{\beta,\gamma}/\pi$. After [8].	127

5.19	Illustration of spectral leakage in (a) Morlet wavelet from 'Airy' family, <i>vs.</i> (b-d) Morse wavelets with $\gamma = 3$, and varying $P_{\beta,\gamma}^2$ values, and their respective Wigner-Ville distributions shown in (e)-(h). After [9].	130
5.20	Panel I: Genuine speech <i>vs.</i> Panel II: Spoofed speech, (a) time-domain signal for the word 'tip', (b) corresponding Morse wavelet-based scalogram, and (c) corresponding low frequency (0 – 40 Hz) scalogram.	131
5.21	Panel I, Panel II, and Panel-III represent the varying distance of the speaker from the microphone, i.e., 5 cm, 5.39 cm, and 6.42 cm, respectively, for (a) time-domain signal for the word 'dad', (b) corresponding scalogram, and (c) selected region of scalogram corresponding to low-frequency (0 – 40 Hz). Solid boxes in red indicate the presence of pop noise. Best viewed in color.	134
5.22	Pop noise energies of various phonemes plotted w.r.t. the distance of the speaker from various microphones for the case, when the speaker is at a distance of 5 cm from Mic 7. The pop noise energies are obtained using the proposed Algorithm 8. The trendlines in each of the sub-figures indicate that the energy of pop noise decreases with the distance of the speaker's mouth from the microphone.	137
5.23	Results (in % Accuracy) for the proposed Morse wavelet-based feature set on the Dev and Eval set of POCO dataset, to observe the effect of wavelet duration parameter (i.e., $P_{\beta,\gamma}^2$).	138
5.24	Results (in % Accuracy) for the proposed Morse wavelet-based feature set on the Dev and Eval set of POCO dataset, in order to observe the effect of γ parameter.	138
5.25	Word-wise VLD accuracies (in %) for the various existing methods compared with the proposed Morse wavelet-based method.	141
5.26	Phoneme-wise analysis of genuine (live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morse wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is present, for (a) plosive (the sample word is 'tip'), (b) fricative (the sample word is 'laugh'), (c) whisper (the sample word is 'who'), (d) affricate (the sample word is 'chip'), (e) nasal (the sample word is 'arm'), and (f) liquid (the sample word is 'run').	142

5.27	Phoneme-wise analysis of spoofed (non-live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morse wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is absent, for (a) plosive (the sample word is ‘tip’), (b) fricative (the sample word is ‘laugh’), (c) whisper (the sample word is ‘who’), (d) affricate (the sample word is ‘chip’), (e) nasal (the sample word is ‘arm’), and (f) liquid (the sample word is ‘run’).	143
6.1	Applications of Voice Privacy.	150
6.2	Discrete-Time Speech Production Model. After [10].	153
6.3	Vocal tract system, $V(z)$ modelled by cascading 2^{nd} order digital resonators. After [10,11].	154
6.4	Mechanical model of differential surface element $d\Sigma$ of vibrating wall, after [10,12].	157
6.5	Glottal and lip boundary conditions as impedance loads. After [10].	158
6.6	Proposed LP-based anonymization system. After [13–15].	163
6.7	%EER (o- original, a- anonymized) for (a) Dev data, (b) test data, and %WER (for two trigram LMs: LM_s -small LM, and LM_l -large LM) for (c) development data, (d) test data (for radius = 0.975 to its value and $\alpha = 0.8$).	165
6.8	Illustration of periodic glottal flow and its spectrum for (a) higher pitch of female speaker, and (b) lower pitch of male speaker. After [10,16].	166
6.9	Panel-I : Analysis of original speech signal. Panel-II: Analysis of anonymized speech signal: (a) spectrogram and speech signal for a female speaker, and (b) spectrogram and speech signal for a male speaker.	167
6.10	Target Selection: By using the Attacker’s ASV to Attack the Victim’s ASV.	168
6.11	Types of speakers based on their vulnerability levels and their effect on EER scores.	170
6.12	Schematic representing effect of Voice Privacy on target selection. .	171
6.13	(a) Attack using target selection, but without voice privacy system, and (b) attack using target selection with voice privacy system. . .	171
6.14	A case study on target selection: EER estimation of each twin-pair.	173
6.15	Publicly available corpora for anti-spoofing research and the associated known attacks. Here, ‘?’ indicates, a gap area to develop anti-spoofing corpora from attacker’s perspective.	175

6.16	Game between an attacker and VP system.	176
6.17	Public key Encryption and Decryption. After [17].	178
7.1	Infant cry modes captured by (a) spectrogram, but have a general structure in (b) CQT-gram, and (c) Morse wavelet-based scalogram representations, due to the form-invariance property followed by CQT and CWT.	186
7.2	Effect of the Morse wavelet parameter $P_{\beta,\gamma}^2$ on Baby Chillanto, and DA-IICT corpora.	188
7.3	Effect of three data augmentation techniques (i.e., tempo, volume, and speed perturbations), on (a) Baby Chillanto, (b) DA-IICT, and (c) combined corpora. Best viewed in color.	188
7.4	Spectrograms of (a) healthy <i>vs.</i> (b) pathological cries.	191
7.5	KLD and JSD of the proposed feature sets. Panel-I and Panel-II denote the cases of non-cepstral features and cepstral features, respectively. KLD (healthy pathology) is shown in (a) and (d). KLD (pathology healthy) is shown (b) and (e). JSD (healthy pathology) is presented in (c) and (f).	195
7.6	Latency period <i>vs.</i> % accuracy between the various non-cepstral features for CQT, u -vector, t -vector, and ω -vector.	196
7.7	Dysarthic speech utterance (for vowel /e/) for male speaker with various dysarthic severity-level (Panel I), corresponding STFT (Panel II), corresponding Mel spectrogram (Panel III), and corresponding Morse wavelet scalogram (Panel IV) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. Best viewed in color.	198
7.8	Scatter plot obtained using LDA for (a) STFT, (b) Mel spectrogram, and (c) Scalogram. Best viewed in color. After [18].	200
B.1	A mass-spring system.	209

List of Tables

2.1	The Best Performing Systems in the Literature for Each of the Anti-Spoofing Corpora.	13
2.2	Results Obtained on ASVSpooF Challenge Datasets for Auditory Transform-Based Systems	15
2.3	Selected Prior Work on VLD	16
2.4	Selected Attacking Techniques in the Literature	20
2.5	Comparison of the Performances of Various SSD Systems w.r.t. Attack Type	25
3.1	Spoofing Algorithms Implemented in the ASVSpooF-2015 Challenge Dataset. After [19].	30
3.2	Details of the ASVSpooF 2015 Dataset. After [20].	31
3.3	Details of the ASVSpooF 2017 v2.0 dataset. After [21].	31
3.4	Statistics of the ASVSpooF 2019 Challenge Dataset. After [22].	33
3.5	Spoofing algorithms used for LA scenario. Here, * indicates neural network-based algorithm. After [22].	33
3.6	Parameter Settings for Replay Configurations in the ASVSpooF 2019 Challenge Dataset. After [23].	34
3.7	Details of speech utterances in the BTAS 2016 database. PH1: Samsung Galaxy S4 phone, PH2: iPhone 3GS, PH3: iPhone 6S, LP: laptop, and HQ is a High-quality speaker. After [24].	35
3.8	Statistics of the ReMASC Dataset w.r.t. Various Acoustic Environments. After [25].	36
3.9	Statistics of the Subset of the ReMASC Dataset Partitioned into Three Subsets. After [25].	36
3.10	Distance of Each Microphone from the Speaker	38
3.11	Statistics of the POCO Dataset Used in This Work. After [26].	38
3.12	Distribution of Words w.r.t. Phoneme Category in POCO Dataset. After [27–29].	39
4.1	Results on the ASVSpooF 2017 v2.0 Database using GMM.	66

4.2	Results on the ASVSpooF 2017 v2.0 Database using CNN.	67
4.3	Results on the ASVSpooF 2017 v2.0 Database using LCNN.	67
4.4	Results on the ASVSpooF 2017 v2.0 Database using ResNet.	68
4.5	Results of Classifier-Level Fusion of the CFCCIF-QESA Feature Set using Different Classifiers on the ASVSpooF 2017 v2.0 Dataset. . . .	69
4.6	Results on Cross-Database Evaluation Between ASVSpooF 2017 v2.0 and ASVSpooF 2019 PA Corpora.	70
4.7	Results on Cross-Database Evaluation Between ASVSpooF 2017 v2.0 and BTAS 2016 Corpora.	71
4.8	Results on Cross-Database Evaluation Between ASVSpooF 2017 v2.0 Corpus and VSDC Corpus.	71
4.9	System Performance When it is Not Under Attack.	72
4.10	System Performance When it is Under Attack.	73
4.11	Results on the ASVSpooF 2019 PA Database using GMM.	73
4.12	Results (in % EER and Accuracy) on the BTAS 2016 Dataset using GMM.	74
4.13	Environment-wise Results (in % EER) Using GMM as the Classifier.	75
4.14	Results in %EER using GMM and CNN as the Classifiers.	75
4.15	Results (in % EER and % AEER) on the ASVSpooF 2015 dataset for various feature sets using GMM as a classifier.	77
4.16	Log Spectral Distance (LSD) Between the LP Residuals of Speech Signal (with $F_s = 16$ kHz) with Various LP Orders (p).	88
4.17	Effect of LP Order on %EER for LFRCC Features on the ASVSpooF 2019 PA Dataset.	89
4.18	Effect of Number of Subband Filters on EER.	89
4.19	Results on Dev and Eval sets for systems trained on GMM, CNN, and LCNN.	97
5.1	Statistics of the POCO Dataset for the Experiments.	107
5.2	Phoneme-wise Average VLD Accuracy (in %).	112
5.3	Average VLD Accuracy (in %) of Different Phoneme Types.	124
5.4	Parameters and the corresponding configurations for replay mech- anism. After [30].	132
5.5	Trendline Equations (in the form of amplitude and time constant (a, b)) Obtained for Each Method w.r.t. Phoneme Type	136
5.6	Results (in % Classification Accuracy) for Morse-CNN-Based Pop Noise Detection Method with Variation in Frequency Range	139
5.7	Phoneme-wise Average VLD Accuracy (in %)	140

5.8	Speaker and Attacker Distance-wise Performance (in % Accuracy) for Morse Wavelet-Based VLD using RC-A (genuine) <i>vs.</i> REP-A (spoof) Dataset with Variations in Subband Frequency Range with CNN as the Classifier.	144
5.9	Overall Performance (in terms of % Accuracy) of Various Feature Sets Across Three Different Classifiers, Namely, CNN, LCNN, and ResNet.	145
5.10	Speaker and Attacker Distance-wise Performance (in % Accuracy) for Morse Wavelet-Based VLD Using RC-A (genuine) <i>vs.</i> REP-A (spoof) Dataset with Variations in Classifier Structure.	146
5.11	Performance Under Ideal Conditions.	146
5.12	Overall performance (in terms of % Accuracy on the Eval set) of the proposed system compared with end-to-end RawNet2 model. . . .	147
6.1	Frequency response of uniform tube with various losses with $p(l, 0) = 0$. After [10,12].	159
6.2	Statistics of the Dev Datasets. After [13].	164
6.3	Statistics of the Eval Datasets. After [13].	164
7.1	Number of utterances in data partitions in each fold.	188
7.2	Overall performance (in % accuracy) of baselines and the proposed features on the three datasets.	189
7.3	% Accuracy for Non-Cepstral and Cepstral u -vector.	193
7.4	% Classification Accuracy for Various Cepstral and Non-Cepstral Feature Set.	193
7.5	% Classification Accuracy of ω -vector with Various Number of Subband Filters.	194
7.6	Results in (% Classification Accuracy) for CNN Classifier.	199
7.7	Confusion Matrix Obtained for STFT, Mel-Spectrogram, and Scalogram.	199

CHAPTER 1

Introduction

Sensitive information needs protection in such a way that only a certain set of people are allowed (authorized) to access it. This authorization to access specific information is given by a biometric system. Biometric systems are used for security purposes in a way that they prevent unauthorized access to important information or data (*information privacy*). The access granted by the biometric is done by capturing traits of humans, which make all human beings unique w.r.t. that particular trait. This thesis focuses on voice-based biometric systems, also known as Automatic Speaker Verification (ASV) systems, given that speech is the most natural and powerful form of communication used by humans to communicate with the outside world. It is the most intuitive, simple, and easy-to-produce characteristic. Furthermore, voice is comparable to other biometrics in many ways. However, voice does have some advantages, not least because the user doesn't need a scanner, such as for iris and fingerprint recognition. Voice is extremely easy to use, and, because of that, it has a higher level of user acceptance than many other biometric identity verification methods. In terms of accuracy, voice is broadly equivalent to other methods, and it is no less secure than fingerprints, retina, or facial recognition. Essentially, voice is both convenient and reliable, not having to be concerned with residues or poor lighting.

ASV systems (as shown in Figure 1.1), deal with verifying speakers' claimed identities with the help of machines [1]. ASV systems have been used for various applications, such as banking transactions and access to systems associated with classified information. In this thesis, the terms voice biometric system and ASV system have been used interchangeably. Since ASV systems have been used for banking transactions and access to buildings associated with classified information, for instance, only the authorized legitimate or *genuine* users are granted access. Nevertheless, some *impostors*, other than zero-effort impostors, deliberately try to fool the ASV system in order to gain an unauthorized access. The deliberate attempts made by the impostor, playing the role of an attacker, are called *at-*

tacks. Besides, due to the recent commercial success of several Intelligent Personal Assistants (IPAs), also known as voice assistants, such as Speech Interpretation and Recognition Interface (SIRI), Amazon Alexa, Google Home, and so on, many voice-enabled devices in Internet of Things (IoT) have also been prone to spoofing attacks [31].

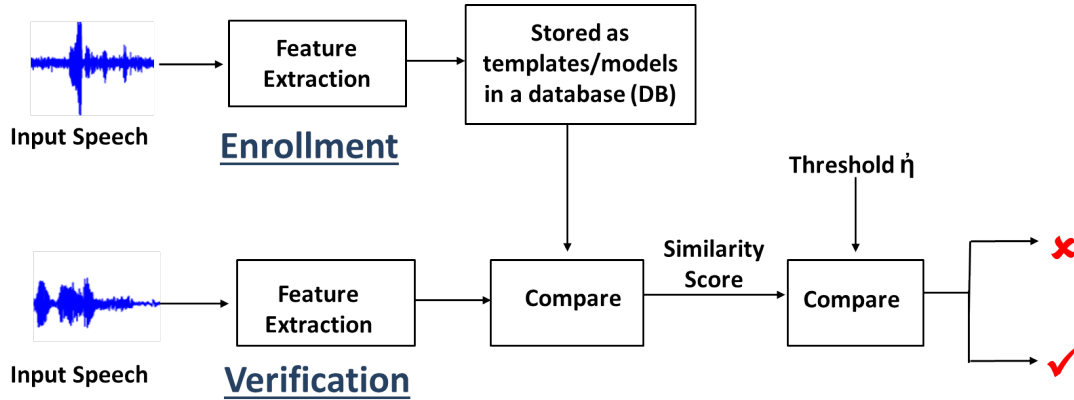


Figure 1.1: A conventional voice biometric (ASV) system. After [1,2].

In this chapter, the motivation and key research objectives of the thesis are discussed. To that effect, we begin with a brief introduction to ASV systems and their vulnerabilities to spoofing attacks in the next subsection.

1.1 Motivation

Spoofing attacks are also called as presentation attacks. Particularly, Voice Conversion (VC) [32,33], Speech Synthesis (SS) [34,35], replay [36–38], twins [39], and impersonation [40] are predominant examples of spoofing attacks on ASV systems [41]. We will discuss some of the spoofing attacks in this subsection.

Impersonation: Modifying one’s voice to sound like a target speaker’s voice consists of an impersonation in which the attacker mimics the characteristic of the target speaker just by using personal skills, without any special technology. It can be performed by professional mimicry artists or, even more so, by identical twins, who exploit behavioral or physiological [40] characteristics, respectively, to sound like the target speaker. Imitation of speech includes adapting high-level features, such as prosody, accent, pronunciation, rhythm, and idiosyncrasies. Such imitation may mislead human perception, however, it is less effective in attacking ASV systems because most of them receive inputs from short-term spectral features to make decisions. Regardless of this fact, we still consider impersonation

as a threat to ASV systems, because spectral features are found to be similar for an identical biological twin-pair [42]. Notably, the characteristics such as fundamental frequency or pitch (F_0) contour, formant contours, and spectrograms are very similar for identical twin-pairs [43]. A real case example of twins fraud occurred in HSBC bank, where a BBC journalist and his non-identical twin spoofed the bank's voice authentication system [44, 45].

Synthetic Speech (SS): Also known as Text-To-Speech (TTS) synthesis, this kind of attack uses text as input to generate speech as output. It mirrors a human speech production mechanism system, i.e., vocal tract and glottal excitation source information, to pose a threat to the ASV system. Due to technological advances, the obtained speech quality sounds considerably natural. Some of the advances which enable this threat are unit selection synthesis [46], statistical parametric synthesis [46], hybrid [47], and Deep Neural Network (DNN)-based methods. Recently, deep learning-based techniques, such as Generative Adversarial Networks (GANs) [48], have been able to perform very well in terms of fooling (or attacking) ASV systems.

Voice Conversion (VC): This attack aims to convert a source speaker's voice to sound like a target speaker's voice [49, 50]. Signal processing techniques, such as vector quantization [51] and frequency warping [52], have been used to achieve successful voice conversion strategies. DNN-based VC utilising wavenets (Wavenet is a deep neural network used to generate raw audio signals from text. It is trained on various Google's TTS datasets, and is well-known to generate natural-sounding speech.) and GANs have received significant attention from the research community.

Replay: It is one of the most convenient attacks to execute but difficult to detect. The attacker uses a pre-recorded speech from the target speaker to get access through the ASV system. Compared with the recorded speech, the genuine data differs only slightly. The differences are due to the impulse response of the recording device and the recording environment. Replay attacks have been a great threat due to the advent of high-quality recording devices, and the careful choice of the recording environment in order to minimize acoustical noise [53–55].

Keeping in mind the vulnerability of ASV systems to so many spoofing attacks, ASVSpooF challenges were organized during INTERSPEECH to boost related research. With ASVSpooF 2015 challenge, various countermeasures (CMs)

were proposed based on different forms of feature extraction techniques on a standard dataset [56–62]. Most of the participant teams focused on signal processing-based research proposals, to design feature sets, and Gaussian Mixture Models (GMMs), as pattern classifiers for the genuine *vs.* spoofed speech detection problem. In addition, multiple CMs for the replay spoof detection were proposed in the ASVspoof 2017 competition [63–65]. In that challenge campaign, there was a paradigm shift from signal processing-based work to sophisticated deep learning approaches. In the ASVspoof 2015 and ASVspoof 2017 challenges, the assessment of CMs was performed independently of ASV systems using Equal Error Rate (EER). Spoofed Speech Detection (SSD) systems were used before the ASV system to detect spoofed speech using CMs, thus making it a two-class problem, as shown in Figure 1.2. However, this type of assessment is not suited for real-world applications involving user authentication. Therefore, in the ASVspoof 2019 challenge, an ASVspoof SSD-centric assessment called *tandem Detection Cost Function* (t-DCF) was proposed to improve the overall reliability of ASV systems, where the SSD system is assessed in tandem with the ASV system.

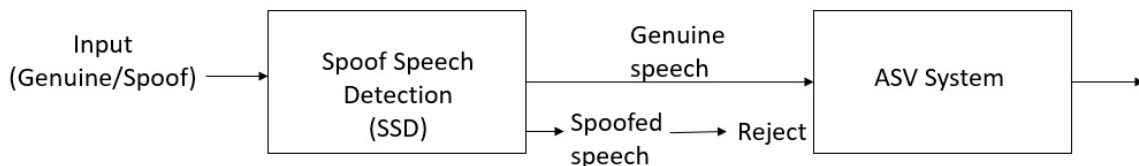


Figure 1.2: Spoofed Speech Detection (SSD) system for ASV system.

1.2 Research Objectives

1.2.1 Replay Spoofed Speech Detection (SSD) Problem

In a replay attack, as shown in Figure 1.3 (a), a recorded version of a genuine speaker’s voice is played back to the ASV system. In particular, the availability of high-quality recording and playback devices has made replay attack detection to be a challenging problem. Moreover, since this type of attack does not require the attacker to have any technical knowledge of the ASV system, or even the speech signal, it is one of the easiest spoofing attacks to be executed.

1.2.2 Voice Liveness Detection (VLD) Problem

In an ideal scenario, an ASV system should be robust against all possible types of attacks. However, unfortunately, in practice, it is not the case. To that effect,

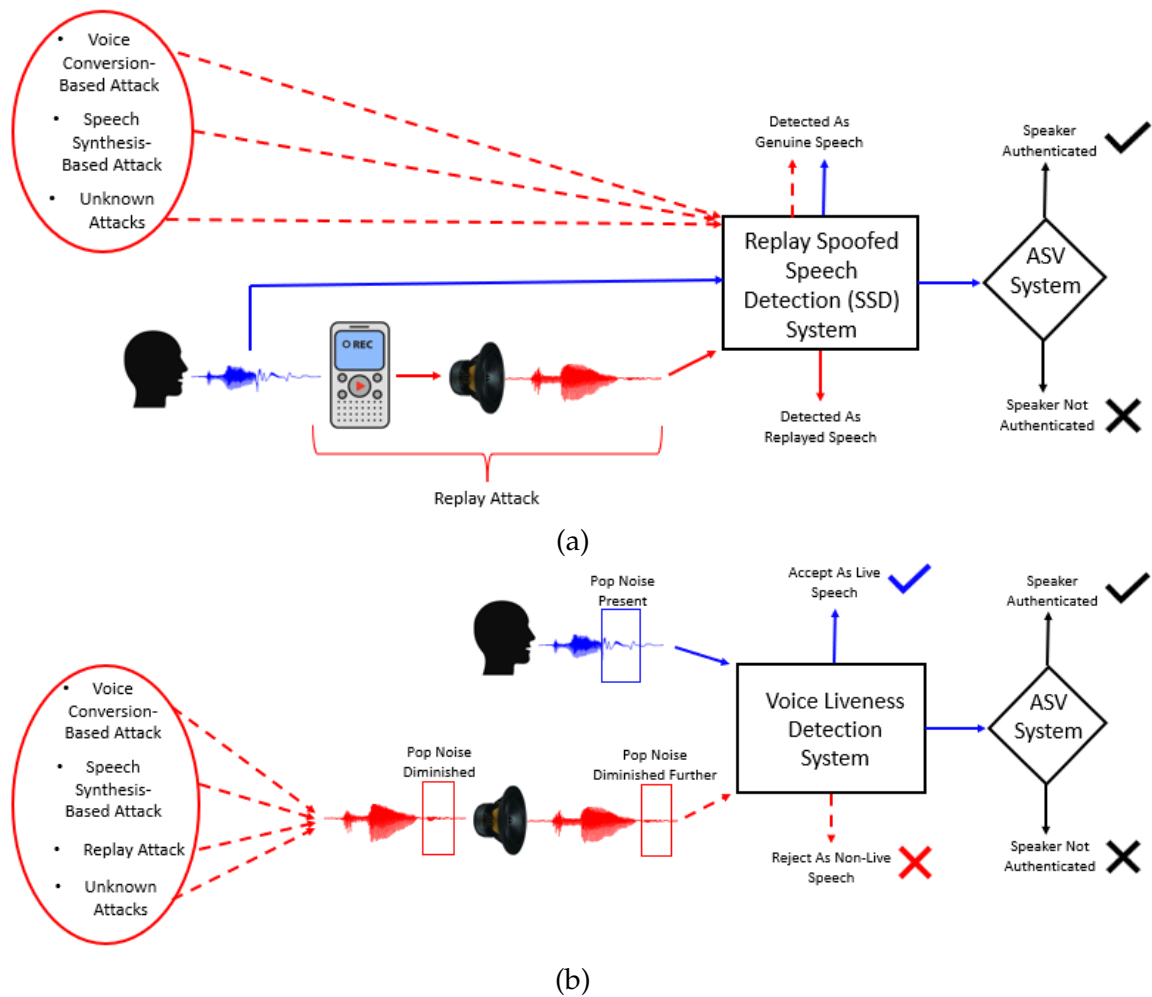


Figure 1.3: Safeguarding an ASV system (a) using a replay SSD system, (b) using a VLD system. Best viewed in color.

defending ASV systems against spoofing attacks is an active research area. One of the defence strategies is to detect whether speech is uttered in a live manner or not. This can be distinguished using VLD systems. To that effect, in this work, we propose to design an efficient VLD system, which detects whether a speech utterance is coming from a ‘live’ speaker or not. The discriminating acoustic cue to detect the liveness is pop noise, which is a characteristic of a live speech, as shown in Figure 1.3 (b). VLD systems can be used to enhance the performance of countermeasure strategies for anti-spoofing systems.

1.3 Contributions and Scope of the Thesis

This thesis is a humble step towards developing defenses against replay spoofed speech. To that effect, two approaches are considered, namely, replay Spoofed

Speech Detection (SSD), and Voice Liveness Detection (VLD).

1.3.1 Features for Replay SSD

- **Quadrature Energy Separation Algorithm (QESA)-based Instantaneous Frequency estimation for Cochlear Cepstral Features (CFCCIF-QESA):** In this work, additional complementary information is captured by introducing a quadrature phase component. To that effect, the existing CFCCIF-ESA feature set is extended to the CFCCIF-QESA feature set, by proposing QESA. Furthermore, IF estimation difficulties are also studied and analysed w.r.t. various IF estimation methods used in the features for replay SSD, including the proposed QESA.
- **Optimized Linear Frequency Residual Cepstral Coefficients (LFRCC):** In this work, optimized LFRCC is proposed w.r.t. replay SSD task. The optimal order for Linear Prediction (LP) is dependent on the sampling frequency of the speech signal. However, this LP order is pertaining to better prediction of speech. However, for the replay SSD task, the goal is not to predict the next speech sample. To that effect, an optimal LP order is found empirically on the ASVSpooof 2019 PA dataset, leading to an optimized LFRCC feature set.
- **Uncertainty Vector:** In this work, Heisenberg's uncertainty principle is exploited to develop a novel feature set known as the *uncertainty vector* (u-vector) for the replay SSD task. For this, the time variance and frequency variance of the signal are used for feature extraction. This analysis is based on second-order moments of the speech signal.

Given that an attacker, in principle, can use any spoofing technique of their choice, it is important to develop countermeasures that are *independent* of the type of attack. To that effect, it is imperative that the countermeasure system should rely on the acoustic characteristics of the genuine (i.e., live) speech, rather than relying on the characteristic of a spoofed signal. To that effect, Voice Liveness Detection (VLD) system comes into play, which relies on *pop noise* as the acoustic characteristics of the live speech. The signal processing-based features for the same are proposed in this thesis, which is briefly discussed next.

1.3.2 Features for Voice Liveness Detection (VLD)

This thesis proposes analytic wavelet-based methods for the VLD task, where the VLD task aims to detect live speech independent of the spoofing attack. To that effect, three analytic wavelets are considered for feature extraction.

- **Bump Wavelet-Based Features:** Bump-wavelets are analytic wavelets, which are exploited for the VLD task in this work. The features pertaining to Bump-wavelet are based on the Continuous Wavelet Transform (CWT), which give a time-frequency representation in terms of a scalogram.
- **Morlet Wavelet-Based Features:** In this work, Morlet wavelets are used for feature extraction for the VLD task. Moreover, Morlet wavelet is closely related to human perception (for both hearing and vision) [66]. To that effect, CWT-based features using Morlet wavelet are proposed.
- **Generalized Morse Wavelet (GMW)-Based Features:** The practical issue of selecting an appropriate wavelet is alleviated to a certain extent because GMWs act as a superfamily of a variety of analytic wavelets [8, 9]. Furthermore, unlike the other (approximate) analytic wavelets, such as the popular Morlet wavelet, Morse wavelets show *strictly* analytic behaviour, i.e., they do not have *spectral leakage* in the negative frequency regions [8]. To that effect, Morse wavelet-based features are also proposed for the VLD task in this thesis.

1.3.3 Voice Privacy and Attacker's Perspective

The assessment of security of ASV systems can be performed whenever various possible approaches and attackers' perspectives are known *a priori*. Hence, possible vulnerability aspects should be examined from the attacker's perspective in order to make an ASV system robust against spoofing attacks. This thesis attempts to study various attacking approaches in the literature. Such a study from the attacker's perspective is also important in order to design robust CM systems. The most crucial information exploited by an ASV system is the speaker's identity (although implicitly). To that effect, the approach of *target selection* is discussed which enables the attacker to select the most vulnerable speaker to target his/her attack, to increase the chances of a successful attack. In particular, an experimental analysis is shown w.r.t. 17 twin-pairs, out of which the most vulnerable twin-pair is selected. However, if privacy preservation is exercised for a speaker's identity,

numerous attacks can be obliterated simultaneously. To that effect, an improved Linear Prediction (LP)-based Voice Privacy (VP) system is also presented.

1.3.4 Additional Applications

Apart from the contributions related to anti-spoofing, the proposed feature sets are evaluated for several Assistive Speech Technologies (AST), such as infant cry classification, and dysarthric severity-level classification. In particular, infant cry classification is done using two approaches, namely, Morse wavelet-based features, and the u-vector. Furthermore, the Morse wavelet-based features are also exploited for dysarthric severity-level classification.

1.4 Organization of the Thesis

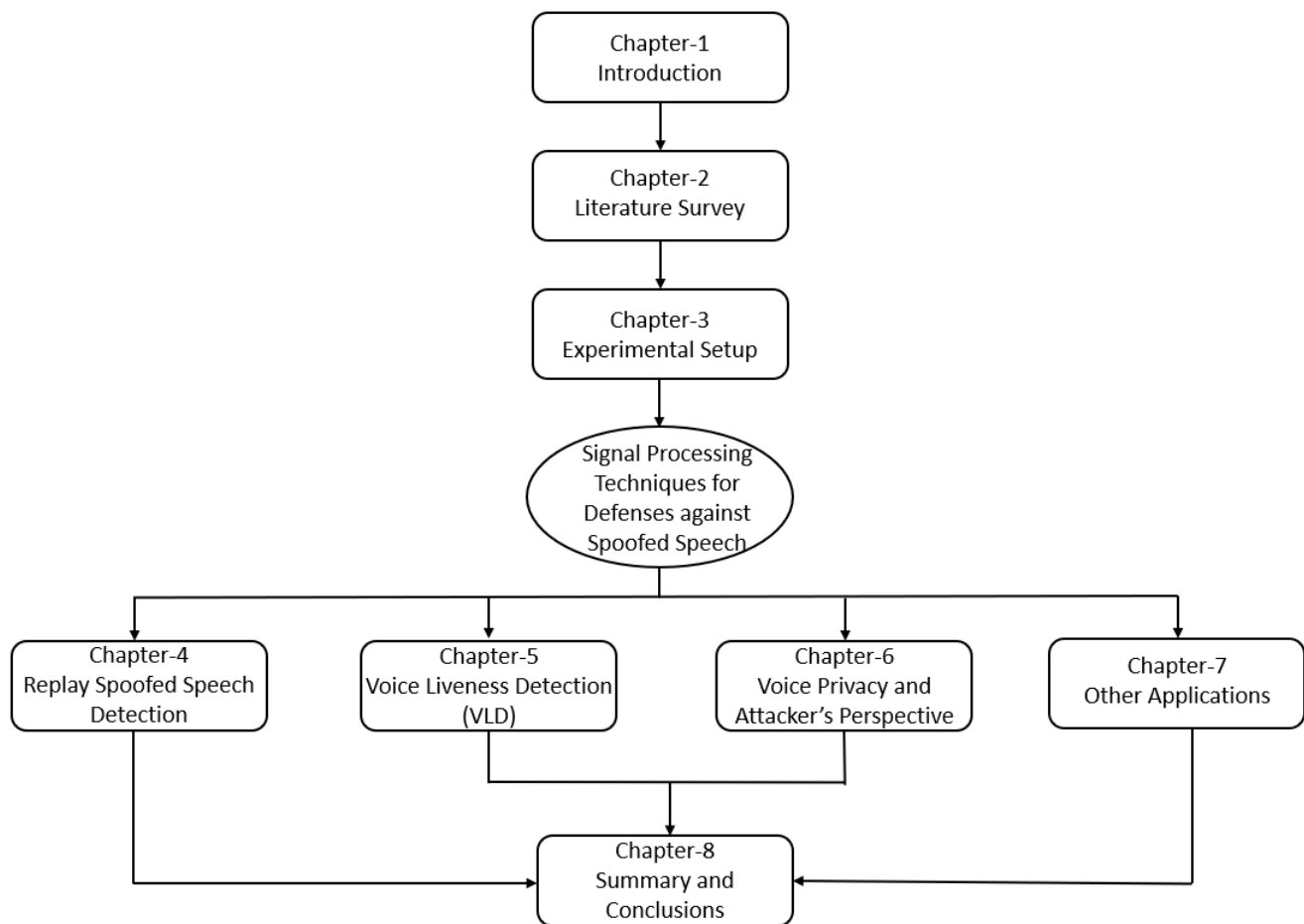


Figure 1.4: Organization of the thesis.

The rest of the chapters in this thesis are organized as shown in Figure 1.4, and details of the organization is briefly described next.

- **Chapter 2** discusses the literature survey on anti-spoofing systems w.r.t. replay SSD, VLD, and the attacker's perspective.
- **Chapter 3** presents details of the experimental setup that is used to perform experiments reported in this thesis. To that effect, the anti-spoofing datasets, classifiers, performance evaluation metrics, and score-level fusion techniques have been discussed in this chapter.
- **Chapter 4** presents the details of three proposed features for replay spoof detection, namely, CFCCIF-QESA, LFRCC, and u-vector. The experimental findings w.r.t. each of these features on the various anti-spoofing datasets are presented in this chapter.
- **Chapter 5** presents the details and the experimental findings of three analytic wavelet-based proposed features for the VLD task, namely, Bump wavelet-based, Morlet wavelet-based, and Generalized Morse Wavelet-based features.
- **Chapter 6** discusses the importance of attacker's perspective for overall security of ASV systems. To that effect, the design of LP-based voice privacy system is presented. Furthermore, the approach of target selection of selecting the most vulnerable speaker from the attacker's perspective is discussed.
- **Chapter 7** includes additional contributions of this thesis for the tasks related to Assistive Speech Technologies (AST), such as classification of normal *vs.* pathological infant cry, and severity-level classification of the dysarthric speech.
- **Chapter 8** presents the summary and limitations of the work presented in this thesis. Furthermore, this chapter also puts forward the open research problems towards replay SSD and VLD tasks.

1.5 Chapter Summary

This chapter covered a brief overview of the ASV system, and its vulnerability to potential spoofing attacks. To that effect, the motivation to design CM solutions to counteract such vulnerabilities is discussed. In this regard, the scientific community's initiative and efforts to comprehensively pose the anti-spoofing challenges were also highlighted. This chapter provided a brief summary of the contributions made by this thesis, focusing on the design of several feature sets for replay

SSD and VLD tasks, along with attacker's perspective and design of voice privacy system. The next chapter presents a literature review on the development of existing CM systems for replay SSD, VLD task, and attacker's perspectives. Furthermore, the research gaps in the literature and the corresponding contributions of this thesis are also included in the next chapter.

CHAPTER 2

Literature Survey

2.1 Introduction

As discussed in Chapter 1, an ASV system is used to grant access to only an enrolled set of speakers (users). All the remaining speakers are treated as non-genuine or imposter speakers or fraudulent attackers. Nevertheless, some impostors deliberately attempt to get unauthorized access to the ASV system. These deliberate attempts made by the impostor (i.e., attacker) are called as spoofing attacks on ASV. This chapter discusses the literature survey on three major aspects of the security of ASV systems, namely, replay SSD, VLD, and the attacker's perspective on ASV systems. Various anti-spoofing measures have been proposed in the literature, which are coherently presented in this chapter. Furthermore, a few research issues have been identified as research gap in the literature and thus, this thesis work attempts to fill this gap.

2.2 Replay Spoofed Speech Detection (SSD)

Some of the possible spoofing attacks on an ASV system are impersonation, twins, voice conversion (VC), speech synthesis (SS), and replay. In order to develop robust countermeasures to detect spoofed speech, the first special session on *Spoofing and Countermeasures for ASV* was organized during INTERSPEECH 2013 [67, 68]. Details of various vulnerabilities on ASV systems and their respective countermeasures (CMs) were presented in [67]. The need for standard datasets, protocols, and performance evaluation metrics in this special session led to the ASVSpooF 2015 Challenge organized during INTERSPEECH 2015 [69]. This challenge focused on developing several CMs against SS and VC spoofs using various kinds of feature extraction algorithms on a standard and statistically meaningful ASVSpooF 2015 dataset [56–61]. The CMs in this challenge were primarily focussed on signal processing-based techniques to develop feature sets, and Gaussian Mixture

Model (GMM) as a pattern classifier for the two-class problem of genuine *vs.* spoof speech detection (SSD). Among the various submissions by the participants, some notable submissions were based on various feature sets, such as Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) (which was the winner system during ASVSpooof 2015 challenge [70]), Linear Frequency Cepstral Coefficients (LFCC) [71], and Constant-Q Cepstral Coefficients (CQCC) [71, 72]. Furthermore, in the ASVSpooof 2017 challenge, the focus was exclusively on real replay SSD [63–65], for which the ASVSpooof 2017 dataset was released [21]. Furthermore, in the ASVSpooof 2019 challenge, the focus was on synthetic or simulated replay (also called Physical Access (PA)), SS and VC-based attacks (called as Logical Access (LA)). Lastly, the most recent challenge is the ASVSpooof 2021 challenge with three tracks, namely, LA, PA, and DeepFake (DF) detection [73]. The DF detection task was introduced in the ASVSpooof 2021 challenge, wherein the attacker synthesizes speech in the voice of a target speaker. DF attacks are usually aimed to harm the reputation of a celebrity, or to spread fake news. In the context of ASV systems, the DF detection task can help to evaluate the robustness of CMs which are used to detect various spoofed and compressed data available online. The details of each of these challenge datasets are discussed in the next chapter. Some notable contributions of the best performing systems on each of the ASVSpooof challenge corpora are shown in Table 2.1.

The recent work reported in [74] shows the best performance on the ASVSpooof 2015 challenge dataset. It is based on multi-level transform (MLT) on the octave power spectra of Constant-Q Transform (CQT). To that effect, Constant-Q Multi-level Coefficients (CMC) features are proposed in this work, which perform better than the existing state-of-the-art anti-spoofing systems on ASVSpooof 2015 challenge dataset. On the ASVSpooof 2017 challenge dataset, the work reported in [75] shows the best and optimal performance of 0% EER. It uses Modified Group Delay Cepstral Coefficients (MGDCC) features with ResNet as the classifier. Furthermore, the work in [76] is the best performing system on the ASVSpooof 2019 LA and PA datasets, with evaluation EER of 1.84% and 0.54%, respectively. It enhances the Light Convolutional Neural Network (LCNN) architecture by introducing angular margin-based softmax activation function. Furthermore, on the ASVSpooof 2021 dataset, the best performing system for LA and DF task is based on media codec data augmentation using LCNN, Residual Neural Network (ResNet), and Long Short Term Memory (LSTM). On the other hand, the best performing system on the PA dataset is based on one-class learning framework using GMM and VAE, giving the performance of 24.25% EER.

Table 2.1: The Best Performing Systems in the Literature for Each of the Anti-Spoofing Corpora.

Dataset	Best Performing SSD System	Performance (in %EER)	
		Dev	Eval
ASVSpooF 2015	CMC [74]	0	0.026
ASVSpooF 2017	MGDCC with ResNet [75]	0	0
ASVSpooF 2019 LA	Angular margin-based softmax activation for LCNN [76]	0	1.84
ASVSpooF 2019 PA	Angular margin-based softmax activation for LCNN [76]	0.0154	0.54
ASVSpooF 2021 LA	Media codec data augmentation with LightCNN, ResNet, LSTM Team T23.	-	1.32
ASVSpooF 2021 PA	One-class learning framework based on GMM and VAE Team T07.	-	24.25
ASVSpooF 2021 DF	Media codec data augmentation with LCNN, ResNet, LSTM Team T23.	-	15.64

CMC: Constant-Q Multi-level Coefficients, MGDCC: Modified Group Delay Cepstral Coefficients, LCNN: Light Convolutional Neural Network, LSTM: Long Short Term Memory, GMM: Gaussian Mixture Model, VAE: Variational Autoencoder, DF: DeepFake.

Out of the various spoofing attacks, replay attacks are the easiest to execute, however, very difficult to detect. In particular, it involves playback of a recorded (genuine) victim speaker’s voice, which is captured using a recording device to get fraudulent access into the system. With the advent of high quality recording and playback devices, replay attacks have become all the more difficult to detect. More so, unlike the other spoofing attacks, such as impersonation, VC, and SS, the attacker does not require having any skills or technical background to mount replay attacks; however, at the same time, they are very difficult to detect.

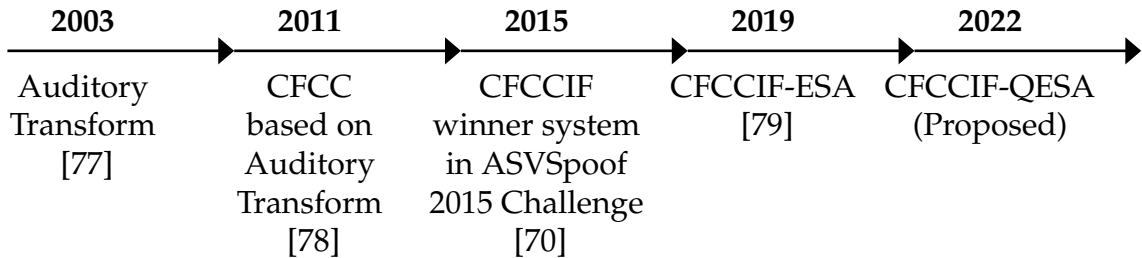


Figure 2.1: A selected chronological progress depicting the development from auditory transform to the cochlear filter-based feature sets.

Next, we discuss the earlier work reported in the literature related to auditory transform-based feature sets (such as CFCC and its variants) for SSD task. Fig. 2.1 depicts the selected chronological progress in the development of auditory transform-based features, which began with the development of auditory (wavelet) transform and its inverse in 2003 [77], where it was validated on real-valued signals in 2009 [80]. In 2011, auditory transform-based features known as CFCC were proposed for a noise-robust speaker identification task [81]. In particular, emphasis was laid on the case when training and testing acoustic environments are mismatched in terms of noise. Here, CFCC features that are developed from the cochlear filter are able to capture robustness in the human hearing system. To that effect, CFCC features have shown remarkable results under mismatched conditions between training and testing [81]. Furthermore, the study reported in [82] used IF spectrum for speech intelligibility. To that effect, subband IF is extracted from the subband filter outputs of CFCC representation for the SSD task [70]. Next, the study in [83] shows that phase-driven characteristics supplement with magnitude-driven characteristics. Similarly, in [70], magnitude characteristics from cochlear filter and analytic or instantaneous phase characteristics via IF were jointly used for SSD task. However, the IF estimation was done by differentiating the analytic phase obtained from Hilbert transform, which is computationally expensive task w.r.t. need for phase unwrapping and also has a poor time resolution as it requires a segment or block of speech data (since Hilbert transform being singular integral in the time-domain requires Fourier-domain implementation and thus, a block of speech data yields better frequency resolution than only a few samples w.r.t. Heisenberg’s uncertainty principle in the framework of signal processing [84]), and detailed proof given in Appendix E. In [79], the estimation of IF was done using Energy Separation Algorithm (ESA), leading to high temporal resolution of IF. To that effect, CFCCIF-ESA feature set was proposed in [79] for the replay SSD task. The performance of each of the cochlear filter-based feature sets is shown in Table 2.2.

Given the success of CFCCIF as winner system for ASVSpooF 2015 challenge [78] and its extension to CFCCIF-ESA for the replay SSD task [79], in this thesis, an improved version of CFCCIF-ESA is proposed, namely, Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm (CFCCIF-QESA), where QESA represents Quadrature-based ESA. Given that the ESA uses a high-resolution, Teager Energy Operator (TEO) for IF estimation, it utilizes only the amplitude information of the three consecutive speech samples. Moreover, due to the absence of Hilbert transform, it does

not utilize the quadrature-phase component of the signal for analytic signal generation. Therefore, in order to incorporate both the advantages, i.e., the excellent time resolution of the TEO and having quadrature-phase component via Hilbert transform, the CFCCIF-QESA feature set is proposed in this thesis as a further improvement in auditory transform-based features. Apart from the high temporal resolution of IF, the proposed feature set exploits the information captured by the *quadrature* relative phase information between the real and imaginary parts of the subband analytic signal. To that effect, the extended definition of TEO for complex-valued signal has been used for the first time for SSD task, resulting in the idea of IF estimation using QESA.

Table 2.2: Results Obtained on ASVSpooof Challenge Datasets for Auditory Transform-Based Systems

Source	Dataset Used	Features Used	Classifier Used	Performance	
				Dev	Eval
[78]	ASVSpooof 2015	CFCCIF	GMM	-	1.211
[21]	ASVSpooof 2017 V2.0	CFCC	GMM	17.60	18.97
[78]	ASVSpooof 2017 V2.0	CFCCIF	GMM	16.61	17.38
[79]	ASVSpooof 2017 V2.0	CFCCIF-ESA	GMM	11.54	14.77

2.3 Voice Liveness Detection (VLD)

Recent research on VLD focuses on pop noise detection. Human breath characterizes live speech because of the ability of microphones to capture pop noise generated from live speech signal [27, 85–88]. Pop noise is a common distortion in speech occurring, when human breath reaches a microphone [27]. It is known to be poorly reproduced by loudspeakers [89, 90]. Therefore, pop noise is a significant discriminatory acoustic cue for VLD.

A selected chronological progress depicting the development of VLD systems is shown in Figure 2.2. With the release of the standard publicly available POp noise COrpus (POCO) recently in 2020, there has been progress towards development of VLD system. It assumes that the live speaker’s mouth is close to the microphone, and signals that are known to spoof ASV systems, such as synthetic speech and replayed speech, fail to reproduce the pop noise as strongly as a live speech signal [26, 100], of course with the assumption that spoofed speech is not recorded with *wiretapping*. Furthermore, Table 2.3 shows the details of the prior work done on the VLD task, where a few of the approaches use their own custom data for the VLD task, due to the unavailability of POCO dataset until 2020. Ap-

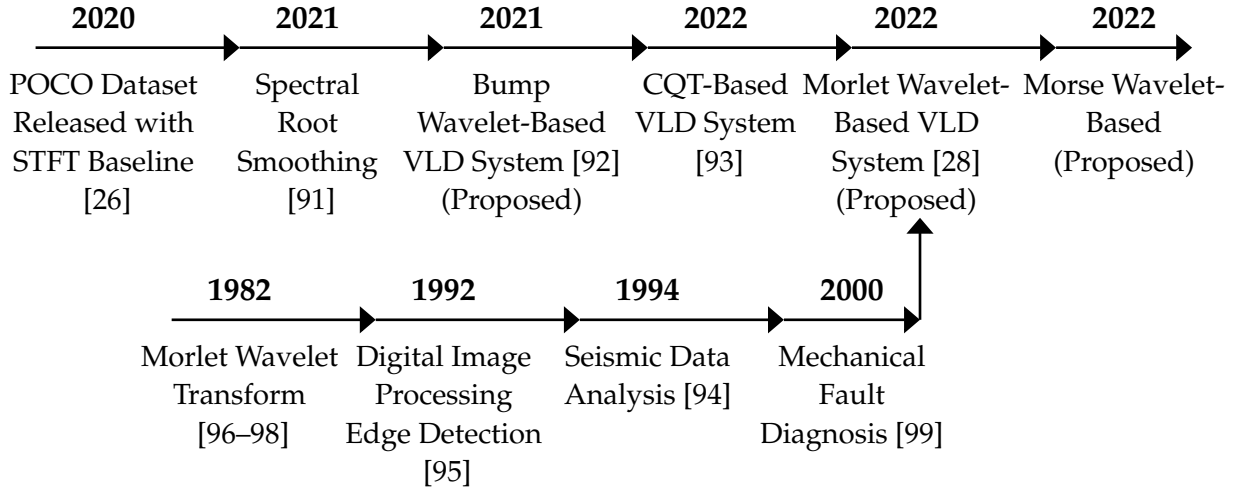


Figure 2.2: A selected chronological progress depicting the development of VLD systems, and the applications of Morlet wavelet in the literature.

Table 2.3: Selected Prior Work on VLD

Source	Dataset Used	Features Used	Classifier Used	Frequency Range	Performance (% Accuracy)
[27]	Custom Japanese Dataset	STFT	GMM	$1 - F_s/2$	5.88 (%EER)
[87]	Custom Data Collected	GFCC	SVM	$1 - F_s/2$	93.5
[101]	Custom Japanese Dataset	STFT	GMM	-	0.95 (% EER)
[26]	POCO	STFT	SVM	1 – 40 Hz	62.08
[102]	POCO	CQT	SVM	$1 - F_s/2$	66.49
[103]	POCO	STFT	CNN	1 – 40 Hz	71.81
[91]	POCO	Spectral Root Smoothing	GMM	$1 - F_s/2$	69.79
[104]	POCO	MGDCC	CNN	$1 - F_s/2$	79.49

STFT: Short-Time Fourier Transform, GFCC: Gammatone Frequency Cepstral Coefficients, CQT: Constant Q-Transform, MGDCC: Modified Group Delay Cepstral Coefficients

proaches, such as low frequency-based single channel detection and subtraction-based pop noise detection have been proposed in [27, 101], on custom dataset which is not publicly available. The feature set used on the custom dataset was the GFCC feature set, which is known to incorporate the characteristics of human peripheral auditory systems [105, 106]. The standard dataset for the VLD task using pop noise detection was released in 2020 as POCO dataset [26]. Thereafter, various approaches on the POCO dataset were proposed in [91, 102–104, 107], which included variations in the type of features used, and the classifiers used. Spectral root smoothing technique with GMM as the classifier was exploited in [91].

However, this approach considered the full-frequency spectrum for the detection of pop noise. Furthermore, the proposition of Bump wavelet-based VLD system in [92] was the first VLD system to exploit wavelet-based technique. Furthermore, the study reported in [93] used Constant-Q Transform (CQT)-based features and was found to perform better than the STFT-baseline.

2.4 Studies from the Attacker’s Perspective

In order to design robust defending mechanisms, it is important to discuss the numerous techniques, which enable spoofing attacks on ASV systems. Assessment of security of ASV systems can be performed whenever various possible approaches and attackers’ perspectives are known *a priori*. Hence, possible vulnerability aspects should be examined in order to make an ASV system robust against spoofing attacks.

Notably, there are two main types of attacks, namely, direct and indirect, as shown in Figure 2.3. Direct attacks are the attacks that are implemented and car-

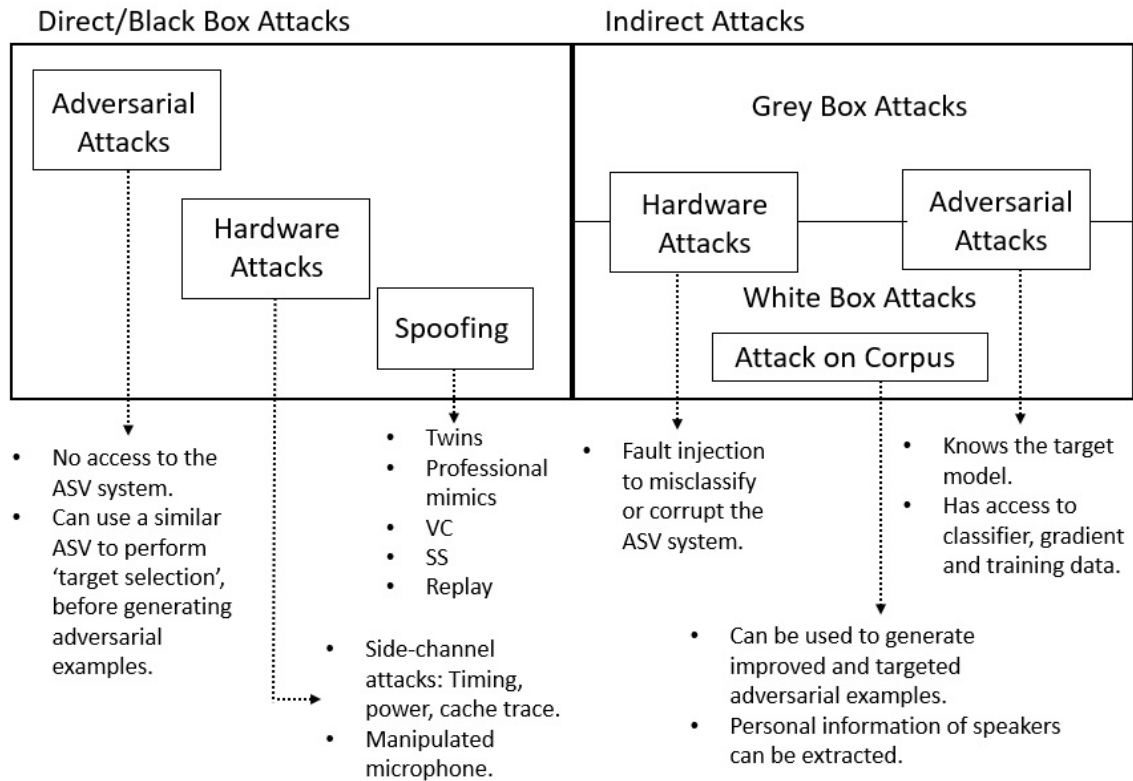


Figure 2.3: Classifying various attacks on an ASV system.

ried out without understanding the internal architecture of ASV system design. As a result, in a direct attack, the attacker does not breach or fool any internal

subsystem in the target ASV system. Instead, attacks are carried out at the microphone and transmission levels. To that effect, a successful direct attack does not need any prior knowledge of the ASV system in question. This is the reason why such an event is also known as *black box attack* [2]. Thus, this kind of attack poses a significant threat to the security of the ASV system due to their ease of execution. Types of direct attacks are spoofing attacks, hardware attacks, and adversarial attacks, as shown in Fig. 2.3.

On the other hand, indirect attacks are those occurring at system-levels, being feasible whenever the attacker has access to the internal subsystems of the target ASV system. If the attacker has complete knowledge and access to all the subsystems, the attack is termed as a *white box* attack. It represents an ideal scenario for attackers, which is not practically realistic. However, despite their unrealistic nature, these attacks should not be ignored since they represent the worst-case possibility for the security of ASV systems. The robustness of an ASV system should be evaluated against such a worst-case scenario so that the ASV systems, and their associated countermeasures, are fully prepared to prevent most of the possible attacks.

A more realistic case of indirect attacks is that in which the attacker has partial knowledge of the target ASV system. Such indirect attacks are termed as *grey box attacks*. Most of the indirect attacks are grey box attacks due to their realistic nature. An attacker can perform more serious damage to the ASV system security by implementing a grey box attack as compared to a black box attack because more power, i.e., knowledge, on the grey-box target ASV system exists.

Spoofing Attacks: Spoofing attacks fall under the category of direct attacks and are the most researched attacks in the ASVspoof literature. Spoofing attacks that are generated from Text-To-Speech (TTS) and Voice Conversion (VC) techniques are called as *Logical Access (LA)* attacks. On the other hand, spoofing utterances which are generated in a real physical space are called as *Physical Access (PA)* attacks. The most common type of PA attack is a replay attack. Furthermore, a recent type of attack, known as *DeepFake* is also a direct attack, which involves generating spoofing utterances using TTS and VC algorithms, similar to LA. Currently, *DeepFake* attacks are known to be the most successful types of attacks. However, the ease of mounting and executing an attack also counts from an attacker's perspective. To that effect, replay attacks are the easiest to mount, and do not even require the attacker to be technically knowledgeable.

Hardware attacks: Due to flaws in hardware implementations of security algorithms, an attacker can find the possibility of mounting a hardware attack. These

attacks can be direct as well as indirect. In case of a direct hardware attack, the attacker can keep track of outputs from the hardware, such as power, timing, and cache traits, to get enough information about the ASV system in order to attack it. Such attacks are called *side-channel attacks*. Simple Power Analysis (SPA) and Differential Power Analysis (DPA), for instance, are classic examples of such types of attacks [108, 109]. On the other hand, in the case of grey box and white box attacks, where the attacker has partial or complete access to the victim’s hardware, the hardware attacks are performed by deliberately mounting faults in the electrical circuitry to alter the behavior of the circuit used. An example of fault injection attack is performed by injecting parametric *Trojan* [110]. With the help of parametric Trojan, the electrical characteristics of the logic gates used in the circuit are altered. However, hardware attacks are usually mounted on systems, which use cryptographic algorithms for their security. In this regard, to the best of the author’s knowledge, a hardware attack on an ASV system is yet to be uncovered, and hence, it is an open research problem.

Attack on corpora: Attacks over *unprotected* corpora are categorized as white box attacks. Attacks over unprotected corpora do not necessarily lead to an attack on an ASV system, however, they can be used to determine personal information about speakers. The ISO/IEC International Standard 24745 on Biometric Information Protection [111] enforces that, for full privacy protection, biometric references should be irreversible and *unlinkable* [112–114]. An unprotected speech corpus, i.e., a biometric reference, enables searching for a speaker’s information on the Internet [40, 115]. Likewise, the study in [116] deals with matching users’ speech to celebrities’ speech data available on YouTube. Thus, due to publicly available speaker data collected from YouTube (also called as “*found data*”), an attacker can look for a celebrity’s voice, which resembles the most to a particular user’s voice, using an approach called as *target selection*, which is described in Chapter 6.

Adversarial attacks: Adversarial attacks aim to intentionally misclassify input data to a Machine Learning (ML) model based on a minor signal perturbation, which forces the ML model to generate an incorrect output. Usually, the perturbation is so modest that it is not even perceivable by humans. The speech signal with the intentionally added perturbation is called as *adversarial example*. An adversarial example w.r.t. to an original speech signal x can be represented as:

$$\bar{x} = x + \delta, \tag{2.1}$$

where δ is so small that \bar{x} is perceptually same as that of x . However, δ is large enough to cause mis-classification. This is in agreement with the finding that there

may exist speech feature parameters that are acoustically relevant for ASV (e.g., fine structure features derived from the open phase region of glottal flow derivative waveform) but perceptually insignificant [10, 117]. Assuming that the ASV system to be attacked is a black box, from the attackers' perspective, to the best of the author's knowledge, the work reported in [118] was the first to propose adversarial attacks against machine learning (ML) methods, where the attacker has no access to a large training dataset. The attack is done by training an attacker's model based on the labels assigned by the existing *victim* ML model. However, the attack presented in this work is not confined to ASV systems, and pertains to more general adversarial attack in machine learning. Notably, in [119], adversarial attacks were evaluated on various scenarios including transferability of attacks, practicability of over-the-air attacks by replay, and human-imperceptibility to demonstrate the imperceptibility of adversarial samples.

With the various types of attacks as discussed, Table 2.4 presents some of the existing attacking techniques in the literature and the respective observations and inferences.

Table 2.4: Selected Attacking Techniques in the Literature

Basis of Attack	Corpus Used	Observations
Choosing the closest target based on FAR using GMM [120]	YOHO	<ul style="list-style-type: none"> • If the number of sessions is more in which the attacker has listened to the target voice, a higher verification error rate is obtained. • The highest FAR achieved was 35% by one of the two imitators.
Choosing the closest target using attacker's ASV on the basis of EER [121]	VoxCeleb1 and VoxCeleb2	<ul style="list-style-type: none"> • Transferability is observed from the attacker's ASV to the attacked ASV in the order of the closest, median, and the farthest speakers. • Contrary to the intuition, if the target speaker's voice is already similar to the impersonator's voice, the verification error is lowered. However, in case of the targets that are not close to the attacker, impersonation increases the verification error, thereby improving the attack.

<p>Training feedback controlled voice conversion system, with feedback coming from the black box target ASV. The VC method used is Phonetic Posteriorgram (PPG)-based [122].</p>	<p>Subset of ASVSpooF 2019 Challenge LA dataset</p>	<ul style="list-style-type: none"> • Higher EER indicates better attack. Overall EER achieved using PPG-VC with feedback attack method was 30.73%, whereas without feedback it was 29.25%. • Male speakers were observed to be more vulnerable due to PPG-based VC attack with EER of 32.90% and 31.60% for the cases of with and without feedback, respectively. • Female speakers on the other hand were comparatively less vulnerable due to reduced EERs of 26.67% and 25%, for the cases of with and without feedback, respectively.
<p>Crafting adversarial examples at the acoustic feature-level, i.e., MFCC and Log Power Magnitude Spectrum (LPMS). To generate perturbation, Fast Gradient Sign Method (FGSM) is used to solve the optimization problem [123].</p>	<p>VoxCeleb1</p>	<ul style="list-style-type: none"> • In black box setting, for perturbation $\epsilon = 20$, EER of 74.62% was achieved. • In white box setting, LPMS <i>i</i>-vector-based system was found to be more vulnerable than MFCC <i>i</i>-vector. For $\epsilon = 1$, FAR and EER obtained by LPMS <i>i</i>-vector were 99.99% and 99.95%, respectively.
<p>Crafting adversarial examples using FGSM and Local Distribution Smoothness (LDS) method [124].</p>	<p>TIMIT</p>	<ul style="list-style-type: none"> • EER is improved by (i) +18.89%, and (ii) +5.54% for <i>the original test set</i> using the regularized model. • Further, EER is improved on the <i>adversarial example test set</i> by (i) +30.11%, and (ii) +22.12%.

<p>Developing an audio-agnostic universal generating sound distortions by estimating perturbation. Robustness is improved by the Room Impulse Response (RIR) [125].</p>	<p>Multi-speaker corpus from Voice Cloning ToolKit (VCTK)</p>	<ul style="list-style-type: none"> • 90% attack success rate is achieved on both x-vector and d-vector-based ASVs. • Attack time is sped up by 100 times. Both were achieved on white box scenarios.
<p>Crafting adversarial examples using Biometrics Transformation Network configuration (ABTN), which jointly processes the loss best of the PAD and ASV systems to generate black box and white box adversarial examples [126].</p>	<p>ASVSpooF 2019</p>	<ul style="list-style-type: none"> • ABTN outperforms adversarial attacks, obtaining 10.28% and 10.14% higher EER joint w. r. t. the PGD ($\epsilon = 1.0$) in the LA and PA test sets, respectively.
<p>Voice conversion using Weighted Frequency Warping (WFW) [127]</p>	<p>TIMIT and CMU-ARCTIC</p>	<ul style="list-style-type: none"> • The WFW-based attack failed on speaker identification systems as the source voice and its corresponding speaker could be identified in numerous cases.
<p>Text-To-Speech (TTS) system, which contains a speaker encoder network, a sequence-to-sequence synthesis network, and an autoregressive WaveNet-based vocoder network, which converts the Mel spectrogram into time-domain signal [128].</p>	<p>VCTK and LibriSpeech</p>	<ul style="list-style-type: none"> • It is demonstrated that synthesized speech is reasonable natural speech, similar to real even on unseen speakers. • Human-level naturalness is not achieved despite the use of a WaveNet vocoder.

An autoencoder-based voice conversion system [129]	VCTK	<ul style="list-style-type: none"> • Generalizes well to unseen speakers. • Speaker characteristics are disentangled from the linguistic content by the encoder bottleneck. • Like [128], it also uses WaveNet vocoder.
SV2TTS [120]	Customized Data	Azure, and WeChat can accept at least 1 synthesized attack utterance.
ASV is trained under unconstrained recording and speaking conditions [130]	Collected Impersonation Dataset (CID)	<ul style="list-style-type: none"> • Attacks using DeepFake speech are more likely to be successful than the other attacking techniques including speech synthesis and impersonation by professionals. • It is established that the fine structures in the speech present due to the human speech production mechanism can capture the discriminating acoustic cues between natural and machine-generated speech, such as DeepFake speech.
DolphinAttack: Inaudible voice commands on ultrasonic carriers [131]	–	<ul style="list-style-type: none"> • Even though inaudible, DolphinAttack voice commands can successfully activate the audio hardware of devices, such as Siri, Alexa, and GoogleNow. • The attack leads to various vulnerabilities, such as visiting a malicious website, spying, injection of fake information, and denial of service.
Targeted adversarial attack called as <i>FAKEBOB</i> under black box setting [119]	LibriSpeech and VoxCeleb	<ul style="list-style-type: none"> • Success rate of 99% is achieved on both open source and commercial systems. • It is concluded that it is difficult to differentiate the speakers of the original voices from those of the generated adversarial voices.

Two attacking setups: Different speaker attack setup, and conversion attack setup [132]	MOBIO and Voxforge	<ul style="list-style-type: none"> • Statistically significant difference with p-value=0 (for males), and p-value=0.0015 (for female) is observed between the mean FAR of the two attacking methods on ASV system. • Conversion attack is significantly more successful than the different speaker attack.
<i>SIRENATTACK</i> : Based on Particle Swarm Optimization (PSO), and fooling gradient method [133]	Common Voice dataset and VCTK	<ul style="list-style-type: none"> • The attack threat is evaluated on the DeepSpeech model, in black box and white box scenarios. • In particular, on ASV systems, average success rate from 91.65% to 99.45% is achieved in black box scenario, on various models.
Professional Swedish Impersonator (male) [134]	–	<ul style="list-style-type: none"> • Low correlation between human perception and speaker verification system is observed. • The human listeners perceive <i>prosodic</i> features in addition to the other speech characteristics. However, machine-based ASV systems do not take prosodic features in account.

2.5 Research Gaps and Contributions of the Thesis

Given that there are numerous methods of attacking an ASV system, such as attacks by twins, professional mimics, VC, SS, and replay, it is important to study the performance of SSD systems for each of these direct attacks. Table 2.5 shows the performance of the SSD w.r.t. each of the attacks. It can be observed that for each attack type, the best-performing system uses a different anti-spoofing method (i.e., varied feature extraction techniques and classifier designs). However, it should be noted that the SSD system performance relies greatly on the attack type, and the dataset. However, considering the practical scenario where an attacker is an external entity who is free to choose any method of generating the spoofed signal, the reliability of SSD systems on a particular attack type makes it far from designing a generalized SSD system.

Table 2.5: Comparison of the Performances of Various SSD Systems w.r.t. Attack Type

Dataset	Attack Type	Best Performing SSD	Performance (in % EER)
Custom Dataset of 17 Twin Pairs	Twins	MFCC with 3 rd order polynomial classifier [43]	97.47*
ASVSpooof 2015	S1, S2, S3, S4	CMC [74]	0
	S5	CAF [135]	0
	S6	CQSPIC-A [136]	0
	S7	CMC	0
	S8	M & P Feats [137]	0
	S9	CMC [74]	0
	S10	CMC [74]	0.221
ASVSpooof 2017 V2.0	Real Replay	MGDCC with ResNet [75]	0
ASVSpooof 2019	LA	CQT, LFCC, FFT, DCT with LCNN [76]	1.84
	PA (Synthetic Replay)	MFCC, CQT, CQCC & VGG, LCNN [138]	0.03
ASVSpooof 2021	LA	Media codec data augmentation with LightCNN, ResNet, LSTM. Team T23.	1.32
	PA (Synthetic Replay)	One-class learning framework based on GMM and variation autoencoder (VAE). Team T07.	24.25
	DF	Media codec data augmentation with LightCNN, ResNet, LSTM. Team T23.	15.64

* denotes the ratio of the number of correctly identified speakers to the total number of speakers [43].

Till now, most of the CMs use acoustical features, such as spectral, F_0 , and modulation-based methods to differentiate an artificial speech signal from natural speech. For each type of spoofing attack, several countermeasures have been proposed. However, no methods have achieved a generalized countermeasure, that can detect *any* type of spoofing attack. From the attacker’s perspective, the attacker is free to mount *any* type of attack on the ASV system. Given the CMs developed so far are *attack-specific*, they may fail to detect if the attack type is different from what the CM is designed for.

For example, as shown in Figure 1.3 (a), a replay SSD is considered, which can be assumed to effectively detect replay utterances. However, if the attacker chooses to deploy VC, SS, or any other unknown methods of attack, the spoofed

signals might go undetected and hence, can be mis-classified as genuine.

However, if we can detect a genuine (live) speech from any other kind of speech (*independent* of the type of spoofing attack), this issue can be alleviated. Interestingly, any kind of speech which is not live, has to be played via a loudspeaker in order to play a spoofed speech. Therefore, if one can distinguish speech produced by a live human being from a speech utterance played via a loudspeaker, live speech can be detected and spoofed utterances can be discarded irrespective of the type of spoofing algorithm or technique used. To that effect, *pop noise* is a distinguishing acoustic cue, which is present in live speech, but is faintly present or absent in speech signals that are played via loudspeakers. Pop noise is generated due to the close proximity of a speaker's mouth with the microphone. In the mechanism of human speech production, the airflow travelling from the lungs to the lips results in a speech wave. If the sound is captured by the microphone at a small distance from the speaker, the microphone captures the speech along with energy released due to the friction between the lips as bursts of airflow, which is termed as *pop noise*. Therefore, pop noise is faint or absent in spoofed signals, more so, in the replayed spoofed signals, because the recording is done discreetly with a large distance from the live speaker.

This thesis work is a humble step to alleviate some of these research gaps in the literature. In particular,

- So far, most of the signal processing-based features have been extracted from the magnitude spectrum of the speech signal. However, the phase (either analytic or Fourier transform phase) characteristics can also be useful for various applications [139–142]. The information captured by the phase has hardly been explored as compared to the magnitude spectrum-based information in the literature [143].

This thesis attempts to address this research gap by proposing quadrature-based ESA (QESA), which is further exploited in proposing the CFCCIF-QESA feature set.

- The earlier studies on replay SSD tasks proposed Instantaneous Frequency (IF)-based features, such as CFCCIF, and CFCCIF-ESA. However, IF estimation suffers from 5 difficulties in the time-frequency literature [144]. Therefore, we believe that the earlier features for SSD (i.e., CFCCIF, and CFCCIF-ESA) also undergo these difficulties.

This thesis attempts to address this research gap by studying and analysing the IF difficulties w.r.t. various IF estimation methods, including the proposed QESA. Such analysis discussed in this thesis led to elimination of a few of the difficulties, whereas the remaining difficulties pose an open research problem.

- The need for a generalized SSD system to alleviate the dependency on the type of spoofing attack, including unknown attacks.

This thesis attempts to address this research gap by proposing analytic wavelet-based methods for the VLD task, where the VLD task aims to detect live speech independent of the spoofing attack.

- Assessment of security of ASV systems can be performed whenever various possible approaches and attackers' perspectives are known *a priori*. Hence, possible vulnerability aspects should be examined from the attacker's perspective in order to make an ASV system robust against spoofing attacks.

- This thesis attempts to study various attacking approaches in the literature. Such a study from the attacker's perspective is also important in order to design robust CM systems.
- The most crucial information exploited by an ASV system is the speaker's identity (although implicitly). If privacy preservation is exercised for a speaker's identity, numerous attacks can be obliterated simultaneously. To that effect, an improved Linear Prediction (LP)-based Voice Privacy (VP) system is also presented.

Apart from these contributions, the proposed feature sets are evaluated for several tasks related to Assistive Speech Technologies (AST), such as infant cry classification, and dysarthric severity-level classification.

2.6 Chapter Summary

This chapter presented the literature survey on voice anti-spoofing for ASV systems, in particular, for replay SSD and VLD tasks. An overview of various SSD approaches along with their performance is discussed in this chapter. Following

this, various attacking approaches are presented briefly, and the need to develop generalized countermeasures is felt as the key research gap in the literature. To that effect, the problem of VLD is introduced, along with literature of existing VLD systems, on the only standard dataset for VLD, known as the POCO dataset. Furthermore, a few research gaps are observed, with the contributions of the thesis in the light to address these gaps. The next chapter discusses the details of the experimental setup that is required for the experimental results and analysis presented in this thesis.

CHAPTER 3

Experimental Setup

3.1 Introduction

This chapter presents the various components of the experimental setup used for the experiments reported in this thesis. To that effect, the elements of the experimental setup discussed are the various speech corpora w.r.t. spoofing detection and voice liveness detection, classifiers used, performance evaluation metrics, and score-level data fusion strategies. This thesis contributes to findings for two tasks w.r.t. anti-spoofing, namely, replay SSD, and voice liveness detection, for which standard and statistically meaningful datasets have been used in this thesis. For spoofing detection, several datasets have been used, such as BTAS 2016, ASVSpooF 2015, ASVSpooF 2017, and ASVSpooF 2019, VSDC, and ReMASC. Out of these, ASVSpooF challenge datasets are aimed to develop CMs for ASV systems, while VSDC and ReMASC datasets are aimed at developing anti-spoofing systems for VAs. For the VLD task, POp Noise detection CORpus (POCO) has been used, which is the only publicly available dataset for VLD research. Furthermore, the classifiers used, such as GMM, CNN, LCNN, and ResNet have been discussed. Finally, the score-level (data) fusion strategies have also been discussed.

3.2 Standard Corpora Used For Anti-Spoofing

As described in Chapter 2, the various ASVSpooF challenges have enabled the progress of research in anti-spoofing w.r.t. various attacks. To that effect, various statistically significant and standard datasets have been released as part of the ASVSpooF challenge campaigns during the years 2015, 2017, 2019, and 2021. ASVSpooF challenges, however, focused on developing CMs for ASV systems. Recently, there has been anti-spoofing research even for VAs and hence, another standard dataset was released, known as the Realistic Replay Attack Microphone Array Speech Corpus (ReMASC). To that effect, the details of all the datasets per-

taining to spoof detection have been discussed in this subsection.

3.2.1 ASVSpooF 2015 Challenge Dataset

The ASVSpooF 2015 Challenge organized during INTERSPEECH 2015 [19], released the publicly available ASVSpooF 2015 challenge dataset to develop CMs against SS- and VC-based spoofing attacks. The dataset consists of 106 speakers, out of which 45 are male and 61 are female speakers. The genuine speech utterances in this dataset are collected in the conditions of minimum background and transmission channel noise. The spoofed utterances, on the other hand, comprise two attacks, namely, SS and VC. To that effect, the dataset in total uses ten algorithms of SS and VC-based methods. These algorithms are denoted from S1 to S10, the details are shown in Table 3.1 [19, 145]. Among these spoofing algorithms, S3, S4, and S10 uses speech synthesis algorithms, and others are VC-based approaches. S1-S9 uses vocoder-based algorithms, and S10 uses unit a selection approach for speech synthesis. Two vocoders, namely, STRAIGHT [146] and Mel log spectrum approximation (MLSA) [147, 148], are utilized to implement vocoder-based algorithm. The details of these spoofing algorithms can be studied in [19].

Table 3.1: Spoofing Algorithms Implemented in the ASVSpooF-2015 Challenge Dataset. After [19].

Subset	# Utterances			Vocoder	Attack Type
	Train	Dev	Eval		
Genuine	3750	3497	9404	None	None
S1, S2	5050	19950	36800	STRAIGHT	VC
S3, S4	5050	19950	36800	STRAIGHT	SS
S5	2525	9975	18400	MLSA	VC
S6-S9	0	0	73600	STRAIGHT	VC
S10	0	0	18400	None	SS

The detailed statistics of the partitions used in the dataset are shown in Table 3.2. The training and development sets of the ASVSpooF 2015 dataset include utterances generated from S1 to S5 algorithms. Since the spoofed utterances generated from S1 to S5 are present in the training set as well as the development (Dev) and evaluation (Eval) sets, they are called *known* attacks. In other words, in known attacks, the training and the testing are done on the same type of utterances generated from the same type of attack. On the other hand, the utterances

generated from S6 to S10 are present only in the Eval set, and absent in the training set. Hence, they are termed as *unknown* attacks.

Table 3.2: Details of the ASVSpooof 2015 Dataset. After [20].

Partition	# Speakers	# Utterances		Spoofing Techniques
		Genuine	Spoof	
Training	25	3750	12625	S1 to S5
Dev	35	3497	49875	S1 to S5
Eval	46	9404	184000	S1 to S10

3.2.2 ASVSpooof 2017 Challenge Dataset

The ASVSpooof 2017 challenge organized during INTERSPEECH 2017, released the publicly available ASVSpooof 2017 dataset to develop CMs against *real* replay spoofs [62]. However, there were anomalies in the dataset, such as the presence of silence regions towards the end of the utterances. To that effect, the second version of the dataset was released known as the ASVSpooof 2017 Version 2.0 dataset [21]. This dataset was primarily introduced for the replay SSD tasks for ASV systems. It contains genuine utterances, and their corresponding replay spoofed speech utterances. It is partitioned into three subsets, namely, training, Dev, and Eval with a brief summary as shown in Table 3.3. The genuine utterances are taken

Table 3.3: Details of the ASVSpooof 2017 v2.0 dataset. After [21].

Partition	# Speakers	# Utterances		Environments
		Genuine	Spoof	
Training	10	1508	1508	E3, E6
Dev	8	760	950	E3, E5, E6
Eval	24	1298	12008	E1-E7

from the RedDots corpus [149]. The replay utterances are generated in 61 different replay configurations for recording devices, playback devices, and acoustic environments. There are a total of 25 recording devices, 26 playback devices, and 26 acoustic environment conditions used in this dataset, with distribution as shown in Table 3.3. The recording devices are categorized into three categories (namely, low, medium, and high), based on their quality and hence, on level of threat posed on the ASV system. Furthermore, similar to the recording devices, the playback devices are also used with varying qualities, i.e., low quality replay devices, such as consumer grade playback devices with a small loudspeaker (i.e., posing low-level of threat), medium quality playback devices with larger loudspeakers (medium threat), and professional audio playback equipment (highest

threat). In addition, the 26 various acoustic environments contribute to the various recording and playback environments. Out of 26, 3 acoustic environments have high ambient noise, thereby posing the least threat. These 3 acoustic environments come from two *balcony* (denoted by E3) and one *canteen* (denoted by E4) environments. Next, the medium-level threat is posed by 18 environments out of which 8 are under *home* conditions (denoted by E5), and 10 are under *office* (denoted by E6) conditions. Lastly, the highest threat is posed by an *echoic room* (denoted by E1), *studio* (denoted by E7), and *analog wire recordings* (denoted by E2). In addition, it can be observed that the replay environments posing the highest level of threat are included in the evaluation (Eval) set, so that the CM models can be assessed on the difficult and unknown attacks.

3.2.3 ASVSpooof 2019 Challenge Dataset

ASVSpooof 2019 challenge organized during INTERSPEECH 2019 [22] released the publicly available ASVSpooof 2019 challenge dataset to develop CMs against Logical Access (LA) and Physical Access (PA) attacks. LA and PA are the two spoofing scenarios in the dataset, which consider the three major spoofing attacks, namely, SS, VC, and replay [22]. The LA scenario has spoofed utterances from SS and VC attacks, while the PA scenario has only replay spoofed utterances. Even though the spoofing attacks considered in the ASVSpooof 2019 dataset are of the same type as in the ASVSpooof 2015 and 2017 datasets, there are still important differences as well. While the ASVSpooof 2015 uses traditional approaches of Text-to-Speech (TTS) and VC, the ASVSpooof 2019 uses neural network-based vocoders for SS and VC-based attacks. Furthermore, while the ASVSpooof 2017 dataset consists of real replayed spoofing utterances, the ASVSpooof 2019 dataset contains simulated replay utterances under *controlled* acoustic conditions, which enable better analysis of results due to the knowledge of simulation parameters to generate simulated replay.

The genuine utterances in the ASVSpooof 2019 dataset are taken from the Voice Cloning Toolkit (VCTK) corpus [150], which has a sampling rate of 96 kHz. For the ASVSpooof 2019 dataset, these utterances are downsampled to 16 kHz with a bit resolution of 16 bits per sample. Table 3.4 shows the details of the various partitions (i.e., training, development (Dev), and evaluation (Eval) sets) of the ASVSpooof 2019 dataset, for LA and PA scenarios, which are discussed in detail further in this subsection.

Table 3.4: Statistics of the ASVSpooof 2019 Challenge Dataset. After [22].

Subset	# Speakers		# Utterances			
	Male	Female	Logical Access (LA)		Physical Access (PA)	
			Bonafide	Spoof	Bonafide	Spoof
Train	8	12	2580	22800	5400	48600
Dev	8	12	2548	22296	5400	24300
Eval	30	37	7355	63882	18090	116640

3.2.3.1 Logical Access (LA)

Logical Access (LA) consists of SS and VC-based attacks. A total of 19 SS and VC algorithms are used, denoted as $A01$ to $A19$, with details as shown in Table 3.5. Out of these, four TTS and two VC algorithms contribute to the utterances in training and Dev sets. Therefore, the spoofed utterances in the Dev set are called *known* attacks. Furthermore, the Eval set comprises spoofed utterances generated from seven TTS and six VC algorithms, where 2 algorithms are used as known attacks, and 11 algorithms are used as unknown attacks.

Table 3.5: Spoofing algorithms used for LA scenario. Here, * indicates neural network-based algorithm. After [22].

Algorithm	Input	Waveform Generator
$A01$	Text	WaveNet* [151]
$A02$	Text	WORLD [152]
$A03$	Text	WORLD
$A04$	Text	Waveform Concatenation
$A05$	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 Concatenation
$A17$	Speech (Human)	Waveform Filtering
$A18$	Speech (Human)	MFCC Vocoder
$A19$	Speech (Human)	Spectral Filtering + OLA

3.2.3.2 Physical Access (PA)

The PA dataset corresponds to the replay attack scenario, wherein an attacker records the genuine speech and replays it in the absence of the genuine speaker, in order to fool the ASV system. The replay utterances in PA scenario are generated using different configurations, such as the attacker-to-talker recording distance (D_a) and loudspeaker quality (Q), whose specifications are shown in Table 3.6. PA is divided into three partitions, namely, training, Dev, and Eval, all of which are disjoint in terms of speakers. Table 3.4 shows the statistics of each of the partitions. The partition of the speakers is similar to that of the LA scenario.

Table 3.6: Parameter Settings for Replay Configurations in the ASVSpooF 2019 Challenge Dataset. After [23].

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

3.2.4 ASVSpooF 2021 Challenge Dataset

The ASVSpooF 2021 challenge dataset was released during a satellite event of INTERSPEECH 2021, with the aim to generalize CMs against LA, PA, and *DeepFake* attacks. This dataset is partitioned in training, Dev, and Eval sets. Out of these, the training and the Dev sets are the same as that of the ASVSpooF 2019 dataset. The ASVSpooF 2021 dataset differs from the ASVSpooF 2019 dataset only in terms of the Eval set, which is partitioned into LA and PA. For LA, the Eval set consists of utterances generated by transmitting genuine utterances through a VoIP network, which leads to the presence of coding and transmission artifacts in the spoofed utterances. However, there is no additive noise. Furthermore, for PA, the Eval set contains real replayed utterances, with a small proportion of simulated replay. These factors are similar to that of ASVSpooF 2019, but are more comprehensive.

The ASVSpooF 2021 challenge introduced DeepFake (DF) detection as the third category of spoofing attack, apart from LA and PA. The DF task aims to detect compressed manipulated speech data of varying characteristics posted online. The scenario is simulated by processing audio files from different sources, along with various codecs used in social media. Therefore, the spoofed utterances are processed with different lossy codecs used typically for media storage, such as *mp3* and *m4a*. The uncompressed data is recovered when the speech is coded and

then decoded, which introduces distortions the speech utterance depending on the type of codec used.

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

This dataset was released during the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2016) [24]. Given that the first ASVSpooof 2015 challenge was focused on spoofing attacks related to *only* SS and VC, the BTAS 2016 dataset was developed, which also includes real replay attack utterances, apart from SS and VC. The replay attack utterances are generated by playback using high quality speakers, laptop speakers, and two smartphones. On the other hand, speech synthesis and voice converted utterances are played with laptop speakers, are also incorporated in this dataset, with details as shown in Table 3.7.

Table 3.7: Details of speech utterances in the BTAS 2016 database. PH1: Samsung Galaxy S4 phone, PH2: iPhone 3GS, PH3: iPhone 6S, LP: laptop, and HQ is a High-quality speaker. After [24].

Data Partition			Training	Dev	Eval
Genuine			4973	4995	5576
Replay	Replay LP LP	R1	700	700	800
	Replay LP HQ LP	R2	700	700	800
	Replay PH1 LP	R3	700	700	800
	Replay PH2 LP	R4	700	700	800
	Replay PH2 PH3	R9	-	-	800
	Replay LP PH2 PH3	R10	-	-	800
SS	SS LP LP	R5	490	490	560
	SS LP HQ LP	R6	490	490	560
VC	VC LP LP	R7	17400	17400	19500
	VC LP HQ LP	R8	17400	17400	19500

3.2.6 Realistic Replay Attack Microphone Array Speech Corpus (ReMASC)

The ReMASC dataset was released to develop CMs for VAs [25]. 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, out of which 22 are female speakers, and 28 are male speakers. Furthermore, out of 50, 36 speakers are native speakers of English language, 12 are Chinese

native speakers, and 2 are Indian speakers. Furthermore, to study the effect of recording device in replay attack, one low quality (iPod Touch (Gen5)), and one high quality recorder (Tascam DR-05) is used. However, it is observed that even with Tascam DR-05, channel and background noise are unavoidable. To that effect, for additional replay source recordings, Google 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 3.8.

Table 3.8: Statistics of the ReMASC Dataset w.r.t. Various Acoustic Environments. After [25].

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

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 3.9. Notably, the partition is *disjoint* in terms of the speakers, and the data distribution among the environments is non-uniform.

Table 3.9: Statistics of the Subset of the ReMASC Dataset Partitioned into Three Subsets. After [25].

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

3.3 Standard Corpora used For Voice Liveness Detection (VLD)

3.3.1 POp noise COrpus (POCO)

The POp noise COrpus (POCO) allows the systematic study of pop noise w.r.t. voice liveness detection [26]. It consists of speech utterances from 66 speakers out of which 34 are female and 32 are male speakers. Significant speakers variation is considered w.r.t. age, English fluency, and accent. The sampling frequency of the utterances in the dataset is 22.050 kHz and the bit rate is 24-bits. There are 3 subsets of the POCO dataset, which are described next.

Genuine utterances with microphone-A (RC-A): For this set of recordings, Audio-Technica AT4040 microphone is used. There is only one microphone in this setting and the distance between the microphone and the speaker is fixed as to be 10 cm, in order to capture the pop noise along with the spoken word. The utterances in RC-A correspond to genuine utterances, as they have pop noise.

Genuine utterances with microphone array (RC-B): This set is another set of genuine utterances. However, it is captured with a microphone array comprising of 15 microphones arranged in a matrix fashion with 5×3 arrangement, as shown in Figure 3.1. The microphones used in this subset are Audio-Technica AT9903

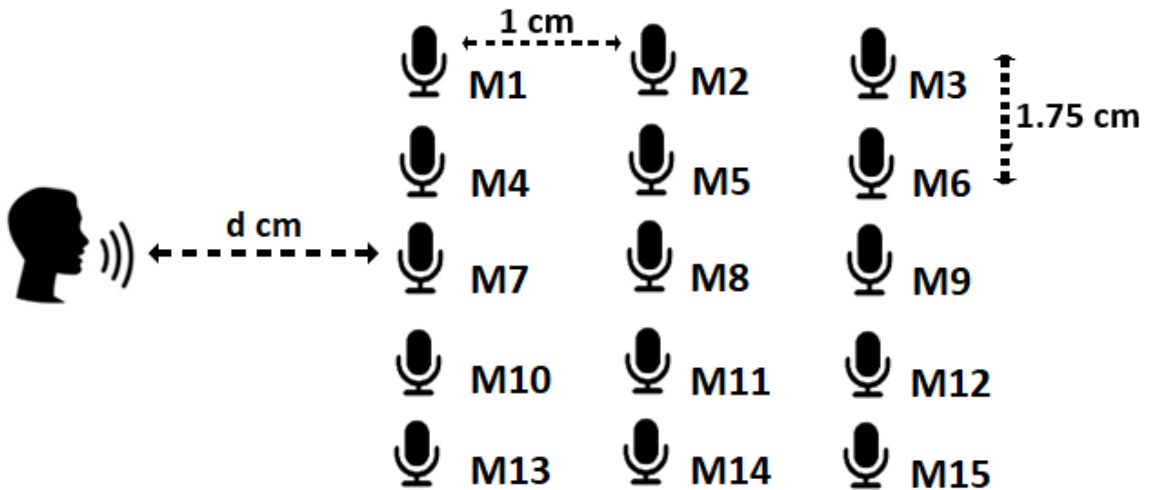


Figure 3.1: The microphone array consists of 15 Audio-Technica AT9903 microphones (M1 to M15) without pop filter. Speaker's mouth is positioned in front of mic M7 at a distance of d cm from the mic M7.

microphones. There are three configurations in this set, each corresponding to a fixed distance between the speaker and the microphone. The 3 distances are 5 cm, 10 cm, and 20 cm. Table 3.10 shows the distance of each microphone from the

Table 3.10: Distance of Each Microphone from the Speaker

Mic ID	Distance Calculation	Distance from the Speaker
M7	d	5 cm
M8	$\sqrt{5^2 + 1^2}$	5.1 cm
M4, M10	$\sqrt{5^2 + 1.75^2}$	5.3 cm
M5, M11	$\sqrt{5^2 + 2.02^2}$	5.39 cm
M9	$\sqrt{5^2 + 2^2}$	5.39 cm
M6, M12	$\sqrt{5^2 + 2.66^2}$	5.66 cm
M1, M13	$\sqrt{5^2 + 3.5^2}$	6.10 cm
M2, M14	$\sqrt{5^2 + 3.63^2}$	6.18 cm
M3, M15	$\sqrt{5^2 + 4.02^2}$	6.42 cm

speaker, when the distance between the speaker and microphone M7 is 5 cm.

Replay utterances with microphone-A (RP-A): For this set of recordings, a single AT4040 microphone is used with a TASCAM TM-AG1 pop filter between the speaker and the microphone. Like the RC-A set, the distance between the speaker and the microphone is fixed to be 10 cm. Given the use of pop filter in this case, this set is emulated and considered to be spoofed and specifically designed for pop noise detection. To that effect, we use additionally embedded reverberation to adapt the replay mechanism. The details of the reverberated POCO dataset are discussed in the next subsection.

Table 3.11: Statistics of the POCO Dataset Used in This Work. After [26].

Subset	# Utterances	# Speakers	
		Male	Female
Training	6952	13	14
Dev	3432	6	7
Eval	6600	13	13

Out of the above mentioned subsets of the POCO dataset, we have used RC-A and RP-A as live and replay utterances, respectively. The speech samples of these 2 subsets were partitioned into training, Dev, and Eval sets, with 40% of the data as the training set, 20% of the data as Dev set, and the remaining 40% of the data as Eval set. Speakers exclusivity is maintained across the partitions. Furthermore, equal distribution of male and female speakers is also considered. The detailed statistics of the partitions is shown in Table 3.11.

There are 44 words in the POCO dataset and their corresponding phonemes in the International Phonetic Alphabet (IPA) have been mentioned in [26]. Given that a word can have multiple phonemes within it, only the most *prominent* phoneme

in the word is taken into consideration. The 44 words of the POCO dataset are put into various phoneme classes as shown in Table 3.12.

Table 3.12: Distribution of Words w.r.t. Phoneme Category in POCO Dataset. After [27–29].

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

3.4 Classifiers Used

In this work, binary classification is done using four types of classifiers, namely, GMM, CNN, LCNN, and ResNet classifiers. While our primary emphasis is on the improved performance due to the various proposed feature sets in this thesis, we also trade it with the effect of different classifiers. The details of each of the classifiers is explained in this subsection.

3.4.1 Gaussian Mixture Model (GMM)

GMM is a parametric model, which is represented as a weighted sum of Gaussian component probability densities. Each component is nothing but a cluster with Gaussian distribution. The probability density function (*pdf*) for a univariate Gaussian distribution is given by [153]:

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

where μ and σ are the mean and standard deviation of a Gaussian distribution, respectively. Similarly, the *pdf* for a multivariate Gaussian distribution is given [153]:

$$f(\mathbf{x}) = \frac{1}{|\Sigma|^2 (2\pi)^{k/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \quad (3.2)$$

where $\mathbf{x} = \{x_1, x_2, x_3, \dots, x_k\}^T$ is a k -dimensional random vector, with mean denoted as $\boldsymbol{\mu}$, and covariance as Σ . Given a particular data point, it has a specific

probability for it to be belonging to each cluster. To estimate the the likelihood of the cluster to which that particular data point belongs, Expectation Maximization (EM) algorithm is used [154]. The EM algorithm gives the Maximum Likelihood Estimation (MLE) parameters. This estimation is done using EM algorithm, which starts with λ (initial model) to estimate $\bar{\lambda}$ (new model), such that $P(V/\bar{\lambda}) \geq P(V/\lambda)$, where V represents a D -dimensional continuous data vector. This estimated new model then becomes the initial model for estimation of the next model. This process is repeated iteratively till it reaches a convergence threshold.

The GMM models the data of the genuine and spoofed speech from the given training speech signals in the form of a statistical model for each class. In the testing (evaluation) phase, the SSD system analyzes the incoming utterance and then estimates the Log-likelihood Ratio (LLR) (expressed mathematically via eq. (3.3)) using pre-trained GMM parameters. In particular,

$$LLR = \log(p(X|\lambda_n)) - \log(p(X|\lambda_s)), \quad (3.3)$$

where $p(X|\lambda_n)$ and $p(X|\lambda_s)$ are the likelihood scores obtained using GMM for natural (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 Convolutional Neural Network (CNN)

CNN is a neural network-based architecture, which consist of one or more convolutional layers followed by classification layers [3, 155, 156], as shown in Figure 3.2. CNNs rely on convolution operation on the input data using a filter (also called as a *kernel*). The kernel is kept smaller than the size of the input and slides over the entire input during its operation. The amount by which the kernel slides is specified by the *stride* size. The convolution reduces the size of the data using the filtering done by the kernel and also reduces the computation cost of the CNN model. In this thesis, features are extracted from the speech signal using various signal processing methods, and fed as input to the CNN classifier. Let the features extracted from the speech signal be denoted by $X \in \mathbf{R}^{f \times t \times c}$, where t , f , and c are the indices of time, frequency, and number of input channels, respectively. The convolution is done using a weight matrix (i.e., kernel) $W \in \mathbf{R}^{m \times m}$, 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. The output of the final convolutional layer is fed to Fully-Connected (FC) layer and the probabilistic output for classification is generated at the output. The Rectified Linear activation Unit (ReLU) function is taken as the

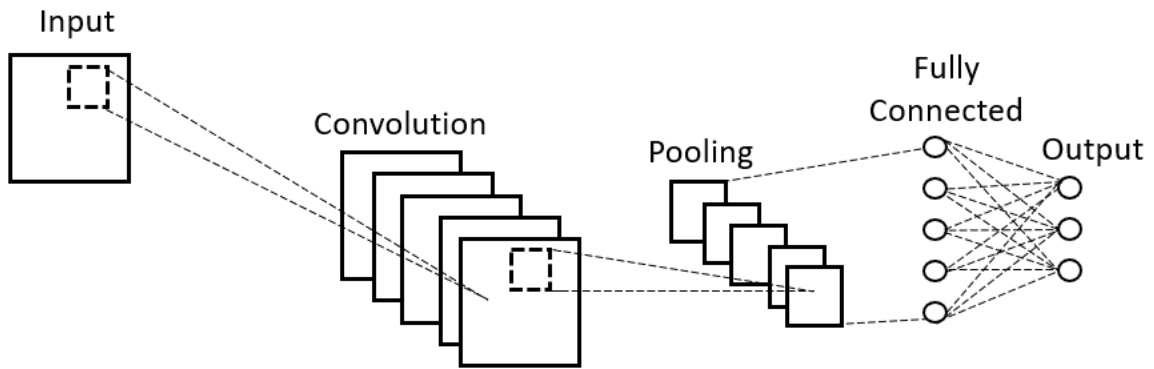


Figure 3.2: Conceptual functional block diagram of CNN. After [3].

activation function for all hidden as well as FC layers [157]. Binary cross-entropy is taken to be the loss function and for optimization of weights, the stochastic gradient descent method is used.

3.4.3 Light Convolutional Neural Network (LCNN)

LCNN is a modified version of CNN, which consists of CNN with a Max-Feature-Map (MFM) activation function. MFM utilizes a competitive selection strategy, which plays the role of efficient feature selection. It is defined as [158]:

$$y_{ij}^k = \max(x_{ij}^k, x_{ij}^{k+\frac{N}{2}}), \text{ where } 1 \leq i \leq H, 1 \leq j \leq W, 1 \leq k \leq N/2 \quad (3.4)$$

Here, x is the input feature vector of size $H \times W \times N$, and y is the output feature vector of size $H \times W \times N/2$. Furthermore, i and j are indices for frequency and time-domain, respectively, and the value of k indicates channel index. MFM suppresses low activation neurons in each layer, and is an alternative to the ReLU activation function. Therefore, it can be considered as a special case of the maxout activation function, which separates noisy data from the rest of the data.

3.4.4 Residual Neural Network (ResNet)

Another deep neural network-based architecture used is ResNet [159]. It integrates the high/mid/low-level features to get the benefit of deeper architectures. DNNs undergo the problem of vanishing/exploding gradients and hence, are unable to learn fine high-level features efficiently. ResNets on the other hand, alleviate this issue by utilizing identity mapping, which allows to stack more layers without introducing the vanishing/exploding gradients and permits the possibility of smooth convergence [159]. As a result, more layers in the architecture

enable one to learn the high-level features efficiently. The ResNet architecture consists of residual layers, followed by FC layers, as shown in Figure 3.3. Here, each residual layer is designed using convolutional layers and ReLU as an activation function. In this thesis, ResNets are used as one of the classifiers due to their success for the SSD task during the ASVSpooF 2019 and ASVSpooF 2021 challenge campaigns [160–165].

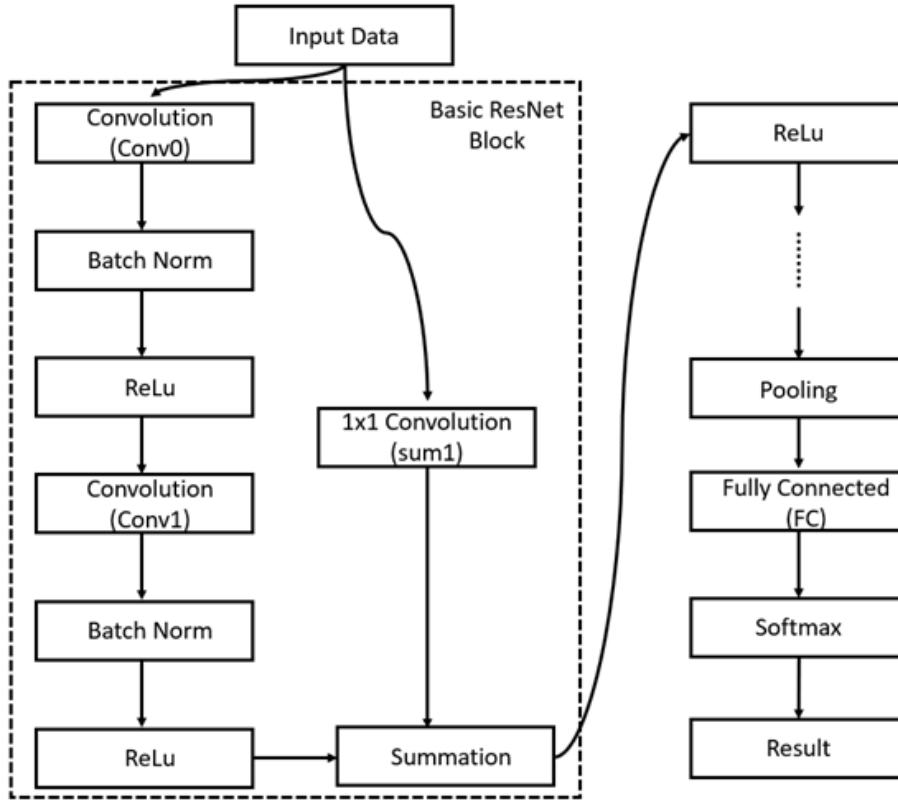


Figure 3.3: Conceptual functional block diagram of ResNet. After [3].

3.5 Performance Metrics Used

The performance metrics used in this work are % Equal Error Rate (EER) and % classification accuracy. As discussed in the subsection 3.4.1, the LLR scores are estimated for testing data using a pre-trained GMM. The LLR scores are used to compute False Rejection Ratio (FRR) and False Acceptance Ratio (FAR). Hence, EER is the point where FRR equals to FAR. Hence, the % EER is given by:

$$\%EER = \frac{FAR + FRR}{2} \times 100. \quad (3.5)$$

For the calculation of % classification accuracy, the first step is to use a classification model, which makes a prediction of class labels for each sample of the

testing dataset. The predicted labels are then compared with actual labels of testing data. The % classification accuracy is then calculated based on the correct prediction of the classification model. The prediction of labels by the classification model is divided into four parts, namely, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The % classification accuracy is calculated from these four parts as [166]:

$$\% \text{ Classification Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \quad (3.6)$$

3.6 Score-Level Data Fusion

The score-level data fusion technique is used on LLR scores, which is evaluated from multiple SSD systems. Score-level fusion enables capturing of the possible complementary information from different SSD systems. The calculation of score-level fusion for two feature sets using *linear weighted sum* is given by:

$$LLR_{fused} = \alpha_i \cdot LLR_{feature1} + (1 - \alpha_i) \cdot LLR_{feature2}, \quad (3.7)$$

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

3.7 Chapter Summary

This chapter presented the details of the key components involved in the experimental setup used in this thesis work. It includes brief discussions on speech corpora, classifiers, performance metrics, and score-level data fusion techniques. The corpora used are divided into two categories - for spoof detection, and for liveness detection. The details pertaining to the data collection, and statistics of the partitions used are given in this chapter. Furthermore, discussions pertaining to the classifiers, evaluation metrics, and data fusion techniques are also included. In the subsequent chapters, several proposed feature sets are discussed w.r.t. the problem of spoof and liveness detection of speech. To that effect, the experimental setup, which is required to validate the performance of the proposed feature sets can be referred from Chapter 3. In subsequent chapters, several proposed feature sets are discussed w.r.t. replay SSD, and VLD task. To that effect, the next Chapter

presents the discussion and results pertaining to replay SSD.

CHAPTER 4

Features for Replay Spoofed Speech Detection

4.1 Introduction

This Chapter ¹ discusses the proposed handcrafted features for the replay SSD task. To that effect, the CFCCIF-QESA feature set is predominantly discussed in this chapter, followed by some additional features, namely, Linear Frequency Residual Cepstral Coefficients (LFRCC) and u-vector, as shown in Figure 4.1. The proposed CFCCIF-QESA feature set is evaluated for the replay SSD task on ASV systems as well as VAs. For ASV systems, experimental results w.r.t. ASVSpooF 2017 V2.0, ASVSpooF 2019 PA, and BTAS are presented in this chapter, where the replay attack scenario in all these datasets is a 1-point replay (1PR). In addi-

¹This Chapter is based on the following publications:

- **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil, "Replay SpooF Detection Using Energy Separation Based Instantaneous Frequency Estimation from Quadrature and In-Phase Components", in *Computer, Speech & Language*, Elsevier, vol. 77 (2023), pp. 101423.
- **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil, "Significance of Quadrature and In-Phase Components for Synthetic Spoofed Speech Detection", in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Chiang Mai, Thailand, pp. 1252-1258, Nov. 07-10, 2022.
- **Priyanka Gupta**, Piyushkumar K. Chodingala and Hemant A. Patil "Energy Separation Based Instantaneous Frequency Estimation from Quadrature and In-Phase Components for Replay SpooF Detection", in *30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, pp. 369-373, 29 Aug. -02 Sep., 2022.
- **Priyanka Gupta**, Piyushkumar K. Chodingala and Hemant A. Patil "Relevance of Quadrature Phase For Replay Detection in Voice Assistants (VAs)" submitted in *31st European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland, Sep. 04-08, 2023.
- **Priyanka Gupta**, and Hemant A. Patil "Linear Frequency Residual Cepstral Features for Replay SpooF Detection on ASVSpooF 2019", in *30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, pp. 349-353, 29 Aug. -02 Sep., 2022.
- Hemant A. Patil, Rajul Acharya, Ankur T, Patil, and **Priyanka Gupta**, "Non-Cepstral Uncertainty Vector for Replay Spoofed Speech Detection", in *30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, pp. 374-378, 29 Aug. -02 Sep., 2022.

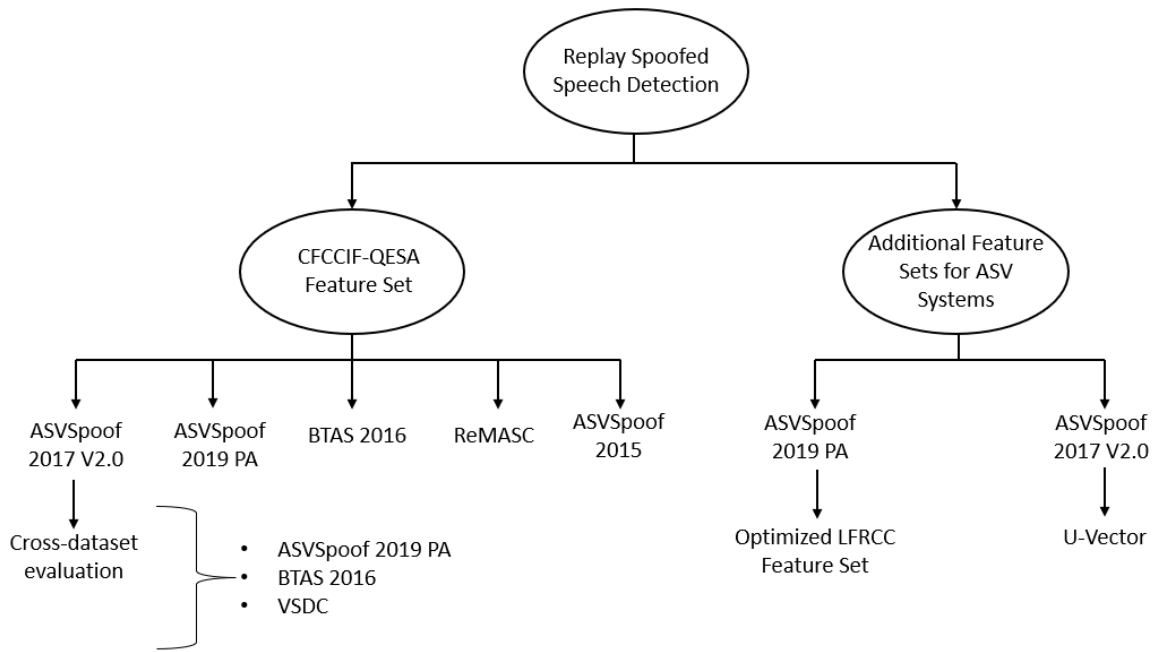


Figure 4.1: Flowchart of the contents of this Chapter w.r.t. the proposed features for the replay SSD task.

tion, the 2-point replay (2PR) scenario is considered by performing cross-dataset experiments using the VSDC dataset. Furthermore, for the sake of completeness, experiments pertaining to the ASVSpooF 2015 dataset are also presented. For VAs, experimental results w.r.t. the ReMASC dataset are presented. The design of the CFCCIF-QESA feature set is explained with a detailed discussion of the various approaches used to estimate Instantaneous Frequency (IF), which is followed by the importance and justification of the choice of quadrature phase (90°) component along with in-phase component by Mutual Information (MI)-based analysis. The incorporation of the quadrature phase enables capturing additional information in the signal, which further improves the performance of the SSD system. In addition to this, a detailed discussion on the difficulties associated with IF is also presented w.r.t. elimination of a few difficulties using the proposed Quadrature-based ESA (QESA) method.

Further, two more feature sets are proposed for the replay SSD task, namely, Optimized Linear Frequency Residual Cepstral Coefficients (LFRCC), and the uncertainty vector (u-vector). In this work, the existing LFRCC feature set is optimized w.r.t. the order of the linear predictor. The development of u-vector is based on the quantification of uncertainty in the form of Time-Bandwidth Product (TBP), which results from the Heisenberg's uncertainty principle in signal processing framework (details given in Appendix E).

4.2 CFCCIF-QESA

4.2.1 Motivation for CFCCIF-QESA

As also discussed briefly by the timeline depicting the development of CFCCIF-QESA in Chapter 2, for the SSD task, in [70], an Auditory Transform (AT)-based Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency (CFCCIF) feature set was proposed. It was based on a cochlear filter and IF-based information. To that effect, IF is estimated conventionally from the analytic phase derived via the Hilbert transform (HT) of the underlying real signal [78]. However, estimating IF from this approach is computationally expensive. Moreover, the resolution of HT in time-domain is poor, as it requires a block (frame) of speech data [167]. To address this issue, the CFCCIF-ESA feature set was proposed [79], which uses the Teager Energy Operator (TEO)-based Energy Separation Algorithm (ESA) [168] to estimate IF with high time resolution for the replay SSD task [169]. Due to the use of TEO in the estimation of IF, CFCCIF-ESA utilizes only the amplitude information of the signal for replay SSD. Moreover, due to the absence of HT, it does not contain the quadrature-phase component of the signal. Therefore, in order to incorporate both the advantages, i.e., the excellent time resolution of TEO and having quadrature-phase component via HT, we propose the CFCCIF-QESA feature set. Here, the term QESA represents Quadrature-based ESA. Furthermore, QESA is based on the extended definition of TEO for complex-valued signals. This extended definition of TEO is exploited for the first time for the SSD task.

4.2.2 Estimation of Instantaneous Frequency (IF)

4.2.2.1 IF Estimation Using Analytic Signal

The IF of a real signal is defined as the time derivative of the unwrapped phase of the analytic signal, whose Fourier transform is zero for negative frequencies [84,144] (details given in Appendix A). The analytic signal $x_a(t)$ corresponding to a real signal $x(t)$ is given by:

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

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

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

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

The use of arctangent function in eq. (4.2) creates a signal processing artifact (due to the periodicity property of arctan) called as *phase wrapping*, thereby creating discontinuities in the phase function, $\phi(t)$. Due to this discontinuity, the IF cannot be derived directly from $\phi(t)$ using eq. (4.3) without the computationally complex task of phase unwrapping [170]. It should be noted that there is no unique definition of IF for a signal, rather there are several difficulties associated with it (to be discussed shortly in subsection 4.2.3.5).

4.2.2.2 IF Estimation Using ESA

The TEO $\Psi\{\cdot\}$ of a continuous-time real signal $x(t)$ is defined as [171]:

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

where $\dot{x}(t)$ denotes the first-order derivative of $x(t)$, and $\ddot{x}(t)$ denotes the second-order derivative of $x(t)$ w.r.t. time t . Furthermore, for a discrete-time signal $x[n]$, the TEO is defined mathematically approximating the derivative operation in eq. (4.4) [171]. In particular,

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

TEO tracks rapid energy (or its running estimate) of the speech signal *within* a glottal cycle with excellent time resolution, requiring only three consecutive samples [169, 171]. Moreover, the TEO enables estimation of the Amplitude Modulation (AM) and Frequency Modulation (FM) components of a speech signal, by the well known ESA, which is briefly described next.

The time-varying amplitude and frequency behaviour in a speech signal is modelled as an AM-FM signal [172] (details given in Appendix C). In particular,

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

where the maximum deviation in frequency is $|q[n]| \leq 1$, $\omega_m \in [0, \omega_c]$, $a(n)$ is instantaneous amplitude, and θ is the constant offset. The instantaneous frequency $\omega[n]$ is given by [10]:

$$\omega[n] = \frac{d}{dn}\phi[n] = \omega_c + \omega_m q[n], \quad (4.7)$$

where ω_c is the carrier frequency. Furthermore, TEO applied on AM-FM signals (such as shown in eq. (4.6)), approximately estimates the product of instantaneous amplitude and instantaneous frequency [172, 173]. In particular,

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

where $\sin^2(\omega[n]) \approx \omega^2[n]$, for $\omega \ll \omega_c$. Thus, it can be observed that both $a[n]$ and $w[n]$ contribute to the running estimate of energy of AM-FM signal representing Simple Harmonic Motion (SHM) [167]. Hence, the following expressions for $a[n]$ and $\omega[n]$ are called as Energy Separation Algorithm (ESA) (detailed proof given in Appendix D) [167]:

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

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

4.2.3 Proposed CFCCIF-QESA Feature Set

This Section shows the analysis and significance of considering the quadrature phase component, which leads to the proposed Quadrature-based ESA (QESA). To that effect, the feature extraction procedure for CFCCIF-QESA is shown in detail in this Section.

4.2.3.1 Optimal Relative Phase using MI

So far, most of the signal processing-based features have been extracted from the magnitude spectrum of the speech signal. However, the phase characteristics can also be useful for various applications [139–142]. The information captured by the phase has not been explored as much as magnitude-based information in the literature [143]. Furthermore, even though phase unwrapping can be avoided by invoking the differentiation property of the Fourier transform (as

shown in [174]), the Fourier transform-based IF estimation fails to explain the IF paradox: "If we have a line spectrum consisting of only a few sharp frequencies, then IF may be continuous and range over an infinite number of values (Chapter 2, pp. 40 [144])". Furthermore, the Fourier transform works on the assumption that each component has a constant frequency at all times. However, in the case of non-stationary signals, such as a speech signal, the frequency is always changing, and frequency is even modulating *within* a pitch period because nonlinear energy modulations take place due to non-linearities in the natural speech production mechanism [175]. Therefore, we cannot depend on the traditional Fourier-based methods, which requires the restriction of the signal being stationary within a period.

Hence, we propose an information-theoretic approach to capture optimal *relative* phase-based information, without estimating phase explicitly. In particular, we exploit the information captured by a signal and its corresponding phase-shifted version. In this context, we employ Mutual Information (MI) as an information theoretic metric to estimate the amount of information between the signal and its corresponding phase-shifted signal. MI analysis is based on information theory (a pivotal work by Shannon [176]), which allows for both the assessment of information quantity within a signal and relationship between different signals. MI of two signals is a measure of dependence of the signals on each other, i.e., a measure of how much information the two signals share [177]. It tells us how much uncertainty about a signal is reduced if we know the other signal. For example, the MI is zero for the case of two signals which are independent, i.e., knowing one of them does not give information about the other signal. Mathematically, MI is estimated as [178]:

$$I(X;Y) = h(X) - h(X|Y). \quad (4.11)$$

Using the joint and marginal probability density functions (*pdfs*) of X and Y , the MI can be also expressed as:

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

Given that the speech signal is modelled mathematically as the sum of several AM-FM signals, we consider AM-FM signal $x[n]$ as [10, 167]:

$$x[n] = [1 + 0.5\cos [60\pi n]] \cos \left[[2\pi f_c n] + 4\sin \left[[2\pi f_c n] + \left(\frac{\pi}{4} \right) \right] \right]. \quad (4.13)$$

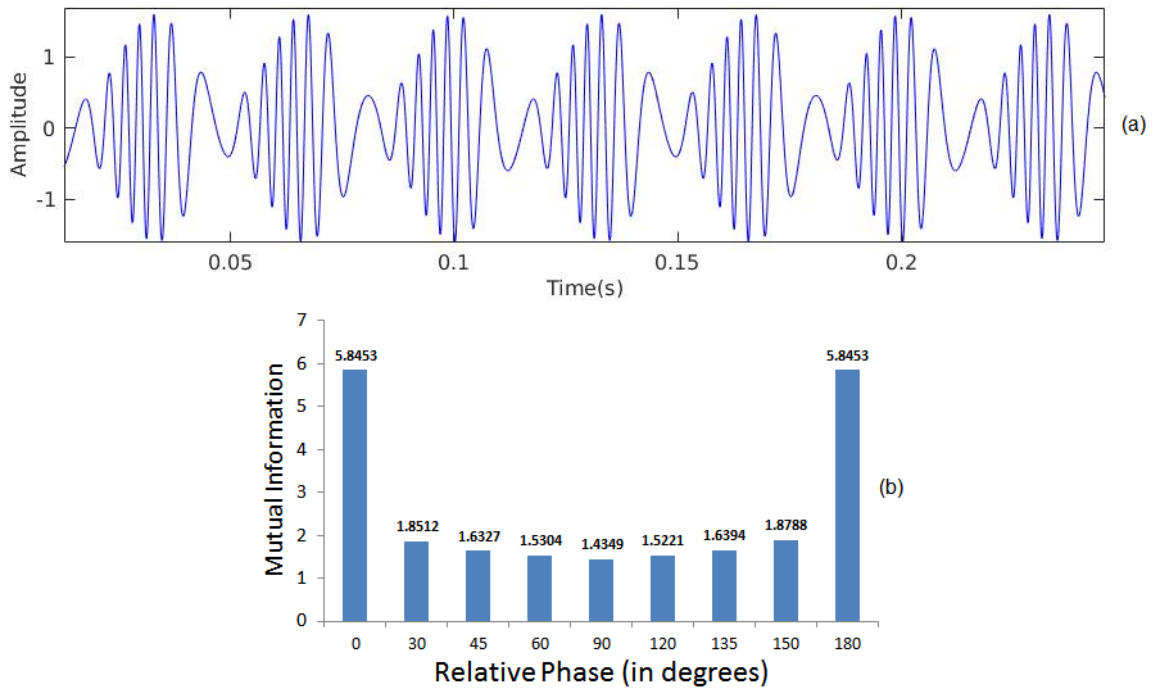


Figure 4.2: (a) AM-FM signal, and (b) MI between AM-FM signal and its phase-shifted version.

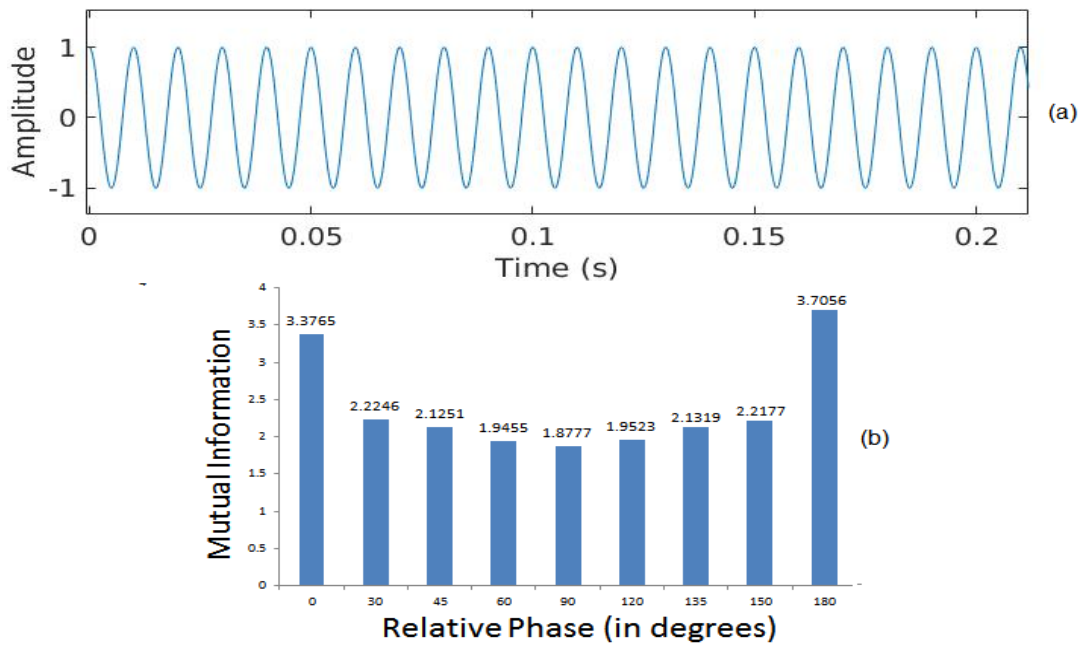


Figure 4.3: (a) Cosine signal, and (b) MI between cosine signal and its phase-shifted version.

We have estimated the MI between $x[n]$ and its phase-shifted versions in order to observe the *optimum* relative phase value. The MI between two time-domain signals is estimated using the Algorithm 1. This algorithm assumes the two input signals X and Y are of equal lengths n . MI is estimated by estimating probabilities

and joint probabilities from histograms of the two signals, using the eq. (4.14).

$$I(X;Y) = h(X) + h(Y) - h(X, Y). \quad (4.14)$$

For constructing the histograms, bin width is selected using the Freedman-Diaconis rule [179], i.e., $bin_width = 2 \frac{IQR(X)}{\sqrt[3]{n}}$, where $IQR(X)$ is the interquartile range of X .

Algorithm 1 MI Estimation of Two Signals X and Y .

- 1: **procedure** MI(X, Y) ▷ X and Y are speech signals of equal lengths
 - 2: $Bin_width1 = 2 \frac{IQR(X)}{\sqrt[3]{n}}$ ▷ $IQR(X)$ means the inter-quartile range of X
 - 3: $Bin_width2 = 2 \frac{IQR(Y)}{\sqrt[3]{n}}$
 - 4: $avg_bins = average(Bin_width1, Bin_width2)$
 - 5: Generate histograms of X and Y using avg_bins
 - 6: Convert histograms to probability values to get $h(X)$ and $h(Y)$
 - 7: Compute joint probability
 - 8: Estimate MI using eq. (4.14)
 - 9: **end procedure**
-

Figure 4.2 (a) shows the AM-FM signal, $x[n]$ modelled by eq. (4.13) and sampling frequency, $F_s = 16$ kHz to generate the corresponding discrete-time signal $x[n]$. This signal is phase-shifted by various angles. MI is estimated between the original AM-FM signal $x[n]$ as shown in eq. (4.13) and its phase-shifted versions. From the MI obtained (shown in Figure 4.2(b)), it can be seen that MI is relatively least (indicating more complementary information), when the optimum phase difference is $\pi/2$. In addition, for a signal $x(t)$, the Fourier transform is computed as [84]:

$$\begin{aligned} \mathcal{F}\{x(t)\} &= \int_{-\infty}^{\infty} x(t)e^{-j\omega t} dt, \\ \therefore X(\omega) &= \int_{-\infty}^{\infty} x(t) \cos(\omega t) dt - j \int_{-\infty}^{\infty} x(t) \sin(\omega t) dt, \end{aligned} \quad (4.15)$$

$$\therefore X(\omega) = X_R(\omega) + jX_I(\omega), \quad (4.16)$$

$$\therefore \angle X(\omega) = \tan^{-1} \left(\frac{X_I(\omega)}{X_R(\omega)} \right). \quad (4.17)$$

From eq. (4.17), it can be observed that the Fourier transform phase $\angle X(\omega)$ is always zero for $X_I(\omega) = 0$ and therefore, if we do not use $\pi/2$ -shifted version of $\cos(\omega t)$ (i.e., $\sin(\omega t)$) as an additional basis function in Fourier transform. In this regard, Figure 4.3 (a) shows the cosine signal and Figure 4.3 (b) shows the

MI obtained between the cosine and its phase-shifted versions. Notably, for the cosine signal as well, MI is observed to be minimum at $\pi/2$ phase shift in $\cos(\omega t)$ (i.e., $\sin(\omega t)$) indicating significance of $\cos(\omega t)$ (i.e., in-phase) and its quadrature component (i.e., $\sin(\omega t)$) in the original definition of the Fourier transform.

To that effect, considering phase-shift as $\pi/2$, we propose an improved relative phase-based CFCCIF-ESA feature set. To incorporate the phase of $\pi/2$, the quadrature component of the speech signal $x[n]$ is estimated using Hilbert transform-based complex-valued analytic signal, $z[n]$. Consequently, TEO for complex-valued signals is used for estimating the signal's energy, and subsequently its IF.

4.2.3.2 Incorporation of Quadrature-Phase Component

MI-based analysis in the previous subsection showed that inclusion of the $\pi/2$ phase gives us additional and complementary information about the signal. In this subsection, we give the description of incorporation of a quadrature-phase component in a bandpass signal. This proposal is shown via a functional block diagram in Figure 4.4.

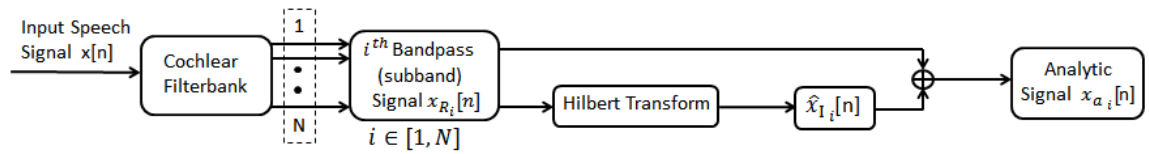


Figure 4.4: Functional block diagram for incorporation of quadrature-phase component. After [4].

The output of each of the subband filters in the cochlear filterbank (used in the proposed CFCCIF-QESA feature set, to be discussed in subsection 4.2.3.6) is a real bandpass signal, $x_{R_i}[n]$ with a cut-off frequency, ω_c [180]. In particular,

$$x_{R_i}[n] = A_i \cos[\omega_c n + \phi[n]], \quad (4.18)$$

where A_i is the amplitude for the i^{th} subband signal. To represent a bandpass signal in complex form, its imaginary part can be represented as:

$$x_{I_i}[n] = A_i \sin[\omega_c n + \phi[n]]. \quad (4.19)$$

It can be observed that eq. (4.18) and eq. (4.19) are in quadrature-phase with each other. In the other words, the Hilbert transform of eq. (4.18) leads us to eq. (4.19). We can say that $x_{R_i}[n]$ and $x_{I_i}[n]$ together represent the time-domain repre-

sensation of the complex bandpass signal. This complex (analytic) representation is known to be a convenient representation of a bandpass signal (Chapter 10, pp. 683-686, [180]). In particular,

$$x_{a_i}[n] = x_{R_i}[n] + j\hat{x}_{I_i}[n], \quad (4.20)$$

where $\hat{x}_{I_i}[n] = \text{Hilbert}\{x_{R_i}[n]\}$.

$$\begin{aligned} x_{a_i}[n] &= A_i[\cos[w_c n + \phi[n]] + j\sin[w_c n + \phi[n]]], \\ x_{a_i}[n] &= A_i e^{j[w_c n + \phi[n]]}. \end{aligned} \quad (4.21)$$

Thus, we can see that $x_{a_i}[n]$ is an analytic signal corresponding to a bandpass signal, $x_{R_i}[n]$.

4.2.3.3 TEO for Complex-Valued Signal

The TEO $\Psi_C\{\cdot\}$ for a complex-valued signal $x(t)$ is given by [181]:

$$\Psi_C\{x(t)\} = \dot{x}(t)\dot{x}^*(t) - \frac{1}{2}[\ddot{x}(t)x^*(t) + x(t)\ddot{x}^*(t)], \quad (4.22)$$

where $*$ denotes complex conjugate operation. When $x(t)$ is real, then the eq.(4.22) reduces to the eq.(4.4), i.e., TEO for a real signal. Furthermore, the complex-valued signal, $x(t)$, can be written in the form of its real and imaginary parts, i.e., $x(t) = x_R(t) + jx_I(t)$. Here, by applying TEO on $x(t)$, we get,

$$\begin{aligned} \Psi_C\{x(t)\} &= \Psi_C[x_R(t) + jx_I(t)], \\ &= \dot{x}_R^2(t) + \dot{x}_I^2(t) - x_R(t)\ddot{x}_R(t) - x_I(t)\ddot{x}_I(t), \\ \Psi_C\{x(t)\} &= [x_R^2(t) - x_R(t)\ddot{x}_R(t)] + [x_I^2(t) - x_I(t)\ddot{x}_I(t)]. \end{aligned} \quad (4.23)$$

Hence, the sum of Teager energies of its real and imaginary parts gives the TEO for the complex-valued signal, i.e.,

$$\Psi_C\{x(t)\} = \Psi_R[x_R(t)] + \Psi_R[x_I(t)], \quad (4.24)$$

where $\Psi_R\{\cdot\}$ is TEO for real-valued signal $x_R(t)$, and can be written as:

$$\Psi_R\{x_R(t)\} = \dot{x}_R^2(t) - x_R(t)\ddot{x}_R(t). \quad (4.25)$$

It should be noted from eq. (4.4) and eq. (4.25) that $\Psi\{\cdot\} = \Psi_R\{\cdot\}$.

Furthermore, considering the eq. (4.22), the term $\dot{x}(t)\dot{x}^*(t)$ will be always real

as it is the product of complex variable \dot{x} and its conjugate. Now, considering the term $[\dot{x}(t)x^*(t) + x(t)\dot{x}^*(t)]$ in the eq. (4.22), let $\dot{x}(t)$ be denoted as the complex number $z_1 = x_1 + jx_2$, and $x(t)$ be denoted as $z_2 = y_1 + jy_2$. Therefore,

$$\begin{aligned}
[\dot{x}(t)x^*(t) + x(t)\dot{x}^*(t)] &= [z_1z_2^* + z_2z_1^*] \\
&= (x_1 + jx_2)(y_1 - jy_2) + (y_1 + jy_2)(x_1 - jx_2) \\
&= [x_1y_1 + j(x_2y_1 - x_1y_2) + x_2y_2] + [x_1y_1 + j(x_1y_2 - x_2y_1) + x_2y_2] \\
&= 2(x_1y_1 + x_2y_2) \in R
\end{aligned} \tag{4.26}$$

Therefore, the extended definition of TEO for complex-valued signals gives real-valued output always.

4.2.3.4 Proposed Quadrature ESA (QESA)

For the i^{th} subband signal $x_{R_i}[n]$ shown in Figure 4.4, IF using ESA is given by [167]:

$$\omega_i[n] = \cos^{-1} \left[1 - \frac{\Psi \{x_{R_i}[n] - x_{R_i}[n-1]\}}{2\Psi \{x_{R_i}[n]\}} \right]. \tag{4.27}$$

In order to incorporate the effect of quadrature phase for the i^{th} subband signal, $x_{R_i}[n]$ is replaced by its analytic signal, $x_{a_i}[n]$ using Hilbert transform as shown in Figure 4.4. The IF in eq. (4.27) is written as:

$$\omega_i[n] = \cos^{-1} \left[1 - \frac{\Psi_C \{x_{a_i}[n] - x_{a_i}[n-1]\}}{2\Psi_C \{x_{a_i}[n]\}} \right], \tag{4.28}$$

where $\Psi_C\{\cdot\}$ is defined as in eq. (4.22). Using eq. (4.20), in eq. (4.28), we get,

$$\omega_i[n] = \cos^{-1} \left[1 - \frac{\Psi \{ (x_{R_i}[n] - x_{R_i}[n-1]) + j(\hat{x}_{a_i}[n] - \hat{x}_{a_i}[n-1]) \}}{2\Psi \{ x_{R_i}[n] + j\hat{x}_{a_i}[n] \}} \right], \tag{4.29}$$

where $\hat{x}_i[n]$ represents the Hilbert transform of $x_i[n]$. Now, using the TEO for complex-valued signals, the eq. (4.29) becomes

$$\omega_i[n] = \cos^{-1} \left[1 - \frac{\Psi \{x_{R_i}[n] - x_{R_i}[n-1]\} + \Psi \{\hat{x}_{a_i}[n] - \hat{x}_{a_i}[n-1]\}}{2 [\Psi \{x_{R_i}[n]\} + \Psi \{\hat{x}_{a_i}[n]\}]} \right]. \tag{4.30}$$

Finally, eq.(4.30) gives us the IF of a complex-valued signal corresponding to i^{th} subband signal, $x_{R_i}[n]$.

4.2.3.5 Alleviation of Some of the difficulties Associated With IF

With respect to time-frequency analysis literature, IF estimation involves 5 difficulties [144], we believe that out of which 2 difficulties are eliminated if the proposed methodology of IF estimation is used. In particular,

1. "IF may be continuous and range over an infinite number of values if there is a line spectrum consisting of only a few sharp frequencies. This means that IF can give frequency values that are not even one of the discrete spectral lines (Chapter 2, pp. 40, [144])."

This happens because IF estimation methods consider the harmonic distortions as continuous intrawave frequency modulations. However, Fourier-based approaches treat the frequency spectral content as discrete harmonic spectral lines. Since TEO (used in ESA for IF estimation) does not use any Fourier-based methods, this paradox is eliminated.

2. "Since IF is instantaneous, the behavior of the signal in the past and future should not be needed. IF gives us the frequencies present in a signal at a particular instant of time. However, to estimate IF using the derivative of the phase of the analytic signal, we have to know the signal for all the times (Chapter 2, pp. 40, [144])."

Given that only 3 consecutive samples are needed to estimate the energy of the signal, TEO is a *nearly instantaneous* operator in the time domain. To that effect, IF estimation using the ESA requires only 5 consecutive samples of the signal [10]. Therefore, this paradox is alleviated to a certain extent.

3. For a band-limited signal, the IF may go outside the frequency band under consideration.
4. IF may not be even one of the frequencies in the spectrum.
If IF is an indication of the frequencies that exist at each instant of time, then how can it not exist in the frequency spectrum?
5. The IF may be negative for negative frequencies even though the spectrum of the analytic signal is zero for negative frequencies (i.e., it is causal in the frequency-domain, details given in Appendix A).

It can be observed that the proposed QESA approach of IF estimation alleviates the first two difficulties associated with IF estimation. To the best of our knowledge and belief, understanding how the proposed QESA alleviates the remaining

three difficulties remains an open research question. To that effect, it should be noted that even the original definition of IF estimation, i.e., $IF = d\phi(t)/dt$ has its own difficulties and which is why there is no unique definition of time-frequency energy density and hence, this topic is difficult and challenging (Chapter 4, pp. 67, [84]).

4.2.3.6 CFCCIF-QESA Feature Extraction

In this subsection, we discuss the computational details of the proposed CFCCIF-QESA feature set, including the selection of *optimal* relative phase, the extended definition of TEO for complex-valued signals, the proposed QESA, and mathematical modelling of the human ear with auditory transform, Basilar Membrane (BM), hair cell representation, and the Nerve Spike Density (NSD).

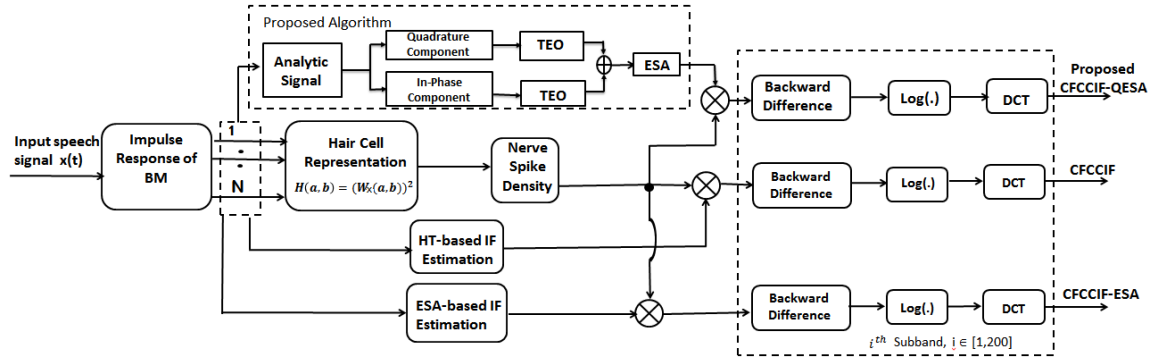


Figure 4.5: Functional block diagram of the proposed CFCCIF-QESA feature set, along with the conventional CFCCIF and CFCCIF-ESA feature sets. The analytic signal in the dotted box is generated w.r.t. procedure shown in Figure 4.4.

The human ear anatomy comprises three main regions, namely, the outer ear, the middle ear, and the inner ear. The visible part of the ear is called as *pinna*, which belongs to the outer ear region. It captures and leads (funnels) the frontal sound waves into the middle ear, which converts acoustic energy of the input sound wave from the outer ear to mechanical energy via the three tiny bones. The fluid in the cochlea is set into motion by the last bone in the ear called as *stapes*. Movement of the stapes puts the fluid inside the cochlea in motion, which further creates traveling waves in the Basilar Membrane (BM). The impulse response of the BM in the cochlea can be represented by a mother wavelet function $\psi(t) \in L^2(\mathbb{R})$ (i.e., Hilbert space of finite energy signals) [84]. The Auditory Transform (AT) models a time-domain speech signal in the cochlea to a set of subband outputs. The AT $Wx(a, b)$, of a real signal $x(t)$ w.r.t. a cochlear filter $\psi(t)$, represents the impulse response of the BM in the cochlear region of the human ear [80].

It is given by [80]:

$$\begin{aligned}
Wx(a, b) &= \langle x(t), \psi_{a,b}(t) \rangle, \\
&= \int_{-\infty}^{+\infty} x(t) \psi_{a,b}^*(t) dt, \\
&= \int_{-\infty}^{+\infty} x(t) \frac{1}{\sqrt{a}} \psi^* \left(\frac{t-b}{a} \right) dt,
\end{aligned} \tag{4.31}$$

where $*$ and $\langle . \rangle$ denote complex conjugate and inner product operation, respectively. Furthermore, a and b are called as the *scaling* and *translation* parameters, and take real values, $a \in R^+$ and $b \in R$. The scaling parameter a is responsible for the dilation and compressing of the mother wavelet function, $\psi(t)$, i.e., it is responsible for the shift in the center frequencies of the subband filters in the cochlear filterbank, resulting in decomposed subband signals. The value of a is given by [80,81]:

$$a = \frac{f_L}{f_c}, \tag{4.32}$$

where f_L and f_c are the lowest and center frequencies of the subband filters in the cochlear filterbank. Thus, a is in the range $0 < a < 1$ when we compress $\psi_{a,b}(t)$ along the time-axis, whereas $a > 1$ when we expand $\psi_{a,b}(t)$. The frequency distribution in the cochlear filterbank can be linear or nonlinear scales, such as Bark, Mel or log [80,81]. In this work, we have used linear frequency scale for frequency distribution. The translation parameter b is responsible for time shift of $\psi(t)$. The factor $\frac{1}{\sqrt{a}}$ in eq. (4.31) ensures that the mother wavelet, and its translated and scaled baby wavelets have equal energies, i.e., $\|\psi(t)\|_2 = \|\psi_{a,b}(t)\|_2 = 1$ [84]. For the AT, the choice of $\psi(t)$ representing the impulse response of cochlear filters is given by:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \left(\frac{t-b}{a} \right)^\alpha e^{-2\pi f_L \beta \left(\frac{t-b}{a} \right)} \times \cos \left(2\pi f_L \left(\frac{t-b}{a} \right) + \theta \right) u(t-b), \tag{4.33}$$

where $\alpha > 0$, $\beta > 0$, $u(t)$ is unit-step function. The value of θ should be selected such that $\int_{-\infty}^{+\infty} \psi(t) dt = 0$, which represents the weak wavelet admissibility condition in the time-domain [84]. In the original study reported in [81], α was chosen to be 3; however, in our study, we found that its optimal value is database-dependent (to be discussed in subsection 4.2.5.8). Similar observations were found for optimization of a parameter β , which controls bandwidth, i.e., quality factor of the cochlear filter [80,81].

Due to the mechanical movements of the BM, the hair cells are displaced in *one*

direction, thereby causing neural activities. The neural excitation stops if the hair cell displacement is in the opposite direction. To that effect, this motion of hair cells is expressed mathematically as [81]:

$$H(a, b) = W_x(a, b)^2; \quad \forall W_x(a, b) \in L^2(R^2), \quad (4.34)$$

where $W_x(a, b)$ is the filterbank output. Furthermore, the output of the hair cell for each subband is converted to a Nerve Spike Density (NSD) count. It is computed by enframing and estimating the average within each frame of length 20 ms with a frame shift of 8 ms (i.e., for the i^{th} subband), and the j^{th} frame number. In particular, NSD is given by:

$$NSD(i, j) = \frac{1}{l} \sum_{b=n}^{n+l-1} H(i, b), \quad n = 1, N, 2N, \dots; \quad \forall i, j, \quad (4.35)$$

where the l denotes the frame length, b is the sample number, and the frame duration is denoted by N . Furthermore, the scales of loudness function are applied on the NSD output [81]. Finally, in order to decorrelate the features, reduce the dimension of the feature vectors, and compaction of energy, Discrete Cosine Transform (DCT) is applied to generate the feature vector.

Algorithm 2 CFCCIF-QESA Feature Vector Extraction.

```

1: procedure CFCCIF-QESA( $x[n]$ )                                ▷  $x[n]$  is the speech signal
2:   for  $i=1:Q$  do                                           ▷ where  $Q$  is number of filters
3:      $f[n] \leftarrow AT(x[n], a[i], b = 0)$                    ▷ AT is the Auditory Transform
     computed using eq. (4.31)
4:      $f_z[n] = f[n] + j \cdot HT\{f[n]\}$                        ▷ Using eq. 4.24
5:      $E_r[n] \leftarrow TEO\{\text{real}(f_z[n])\}$ 
6:      $E_i[n] \leftarrow TEO\{\text{imag}(f_z[n])\}$ 
7:      $\psi\{f_z[n]\} = E_r[n] + E_i[n]$                          ▷ Extended definition of TEO
8:      $IF \leftarrow \text{Cos}^{-1}\left[\frac{1 - \psi\{f_z[n] - f_z[n] - 1\}}{2\psi\{f_z[n]\}}\right]$ 
9:      $c \leftarrow$  Hair Cell Representation ( $f[n]$ )           ▷ Using eq. (4.34)
10:     $d \leftarrow$  NSD ( $c$ )                                    ▷ Using eq. (4.35)
11:     $v \leftarrow d + IF$ 
12:  end for
13:   $feat \leftarrow \log(\text{DCT}(v))$ 
14: end procedure

```

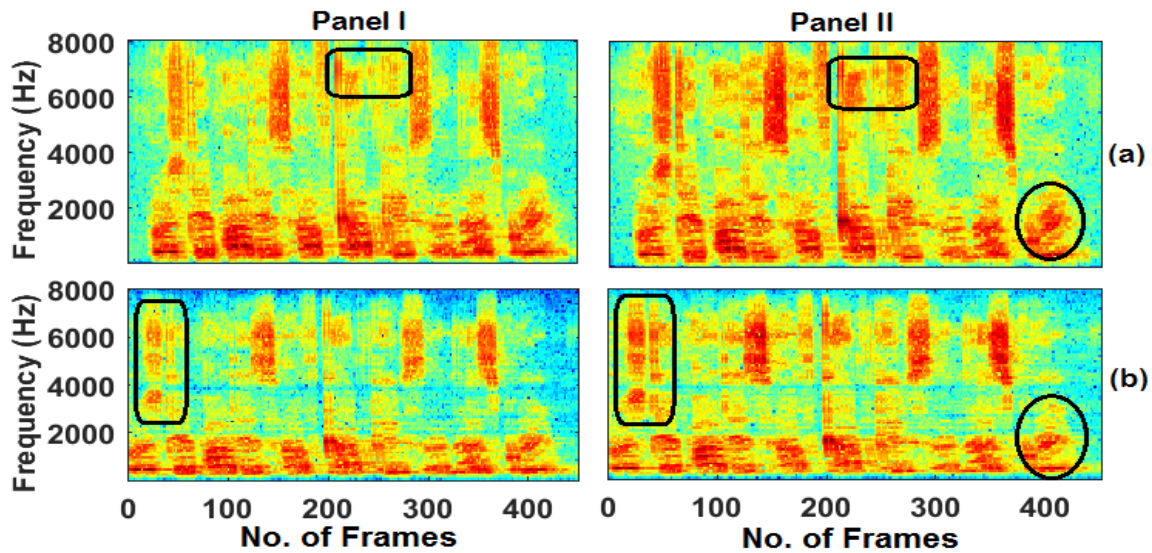


Figure 4.6: Spectrographic representation of the genuine *vs.* spoofed speech. Panel I and Panel II represent the spectrographic representation of CFCCIF-ESA and CFCCIF-QESA, respectively. Here, (a) genuine speech signal, and (b) corresponding spoofed (replay) speech signal.

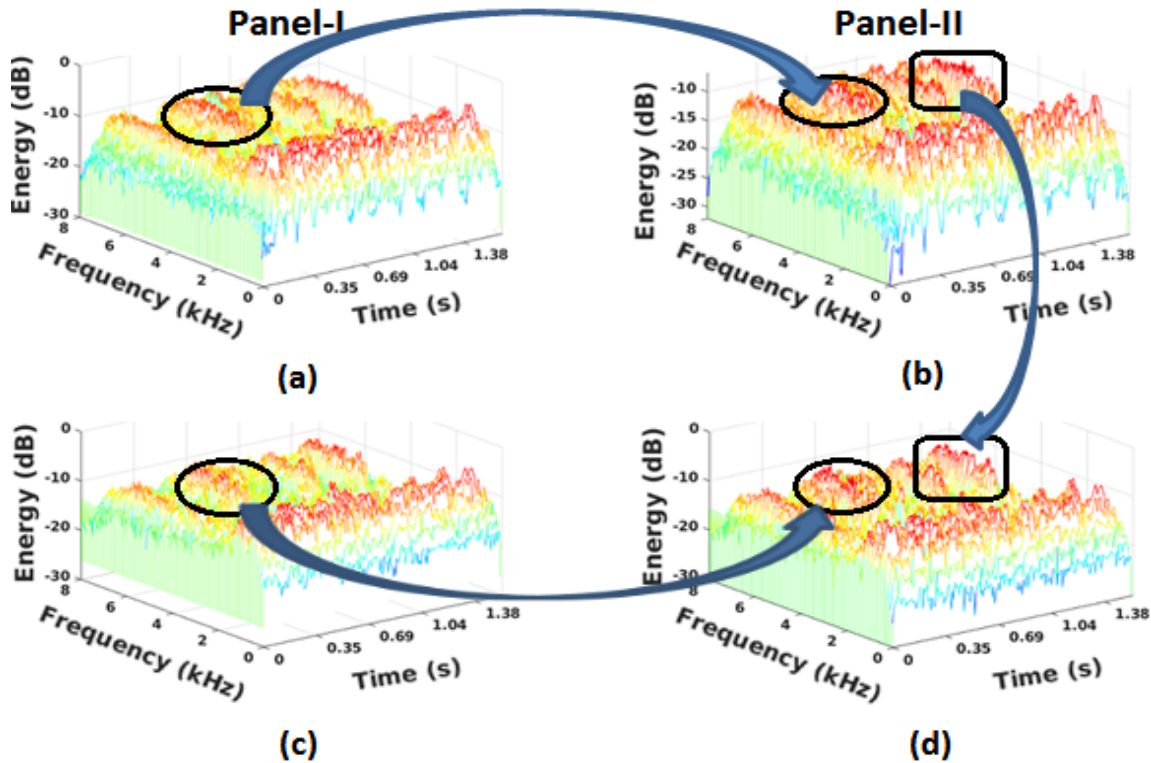


Figure 4.7: Panel I and Panel II represent the analysis of waterfall plots for CFCCIF-ESA and CFCCIF-QESA, respectively, for the same utterances used in Figure 4.6. Here, (a) and (b) represent the waterfall plots for genuine speech signal, and (c) and (d) represent the waterfall plots for spoofed speech signal.

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

To analyze the effectiveness of the CFCCIF-QESA feature set as compared to the CFCCIF-ESA feature set, we observe the spectrograms of both the feature sets for the same speech utterance, which is shown in Figure 4.6. Here, the spectrogram of a feature set represents the Energy Spectral Density (ESD) in the time-frequency plane for that feature set. However, this spectral representation is obtained before the DCT operation during the feature extraction. In Figure 4.6, Panel I and Panel II show the ESD for CFCCIF-ESA and CFCCIF-QESA, respectively. Figure 4.6(a) and Figure 4.6(b) represent a genuine speech utterance and corresponding spoofed (replay) speech utterance, respectively. It can be observed from Figure 4.6 that in the higher frequency region, the resolution of the CFCCIF-QESA feature set is high for both genuine and spoofed as compared to the CFCCIF-ESA feature set. It is highlighted in Figure 4.6 by curved rectangle boxes. Furthermore, the oval boxes show the difference in ESD of genuine *vs.* spoofed speech, for the CFCCIF-QESA feature set. Thus, we believe that due to the presence of the quadrature-phase component, the CFCCIF-QESA feature set captures more dominant spectral energy density as compared to the CFCCIF-ESA feature set, because all the other parameters and conditions for both the feature sets are kept identical.

Figure 4.7 shows the waterfall plots for the CFCCIF-ESA and CFCCIF-QESA feature sets in three dimensions (for the same utterances used in Figure 4.6), demonstrating the competence of the CFCCIF-QESA feature set against CFCCIF-ESA for replay SSD. Panel-I corresponds to CFCCIF-ESA and Panel-II corresponds to CFCCIF-QESA. The waterfall plot enables us to observe the density of bumpy structures of ESD. Comparing Panel-I with Panel-II, it can be observed that CFCCIF-QESA has the capability to capture information more efficiently as compared to the CFCCIF-ESA.

4.2.4 Setup

- **Datasets Used:** The datasets used for the experiments shown in subsection 4.2.5 are ASVSpooF 2015, ASVSpooF 2017 v2.0, ASVSpooF 2019 PA, VSDC, BTAS 2016, and ReMASC. The details of each of these datasets is given in subsection 3.3 of Chapter 3.
- **Classifiers Used:**
 - **CNN:** In this work, the input feature size for CNN is taken to be 30×400 . Since the raw waveform can have varying duration, the input size

is fixed by padding the input with its initial samples till its size becomes 30×400 . The CNN architecture used consists of five convolutional layers (Conv1, Conv2, Conv3, Conv4, and Conv5) followed by two fully-connected layers (FC1 and FC2). Here, in the first two convolutional layers, the data is convolved using a kernel size of 5×5 with a stride of 1 and padding of 2. Furthermore, in the remaining three convolutional layers, the kernel is used of size 3×3 with the stride and padding of 1. Here, after every convolutional layer, max-pool layer is used, having kernel of size 2×2 with a stride of 2, in order to reduce the size of data and also to reduce the computation cost of CNN model. After extraction of features from convolutional layers, the output of Conv5 is fed to the FC1 layer and the probabilistic output for classification is taken from FC2. The Rectified Linear activation Unit (ReLU) function is taken as the activation function for all the hidden as well as FC layers [157]. Binary cross-entropy is taken to be the loss function and for optimization of weights, stochastic gradient descent is used as the optimizer.

- **LCNN:** For the LCNN model, the input feature is of size 30×400 . The varying durations of the input speech waveform are made constant to 30×400 by padding the input with its initial samples till its size becomes 30×400 . The LCNN model consists of four CNN layers (Conv1, Conv2, Conv3, and Conv4) and two FC layers (FC1, FC2). In the convolutional layers, the data are convolved using a kernel of size 3×3 with a stride of 1 and padding of 1. After each layer, the MFM and max-pooling layer is used. The MFM layer uses a kernel of size 3×3 with stride of 1 and padding of 2. The max-pooling is used with kernel size of 2×2 and stride of 2, to reduce the size of feature vector and also to reduce the complexity of the model. The ReLU activation function is used in the FC7 layer. For calculation of loss, we have used binary cross-entropy as loss function stochastic gradient descent as the optimizer.
- **ResNet:** The ResNet architecture used consists of four residual layers (Res1, Res2, Res3, and Res4) followed by two fully-connected layers (FC1 and FC2). Here, each residual layer is designed with two convolutional layers and ReLU as activation function. Furthermore, the Res1 layer has kernel size of 7×7 and stride of 1, while the other three residual layers are used with kernel size of 3×3 and a stride of 1.

4.2.5 Experimental Results

This Section presents the results obtained on various datasets for different evaluation factors. In particular, first, the experiments are performed to optimize the parameters of the proposed CFCCIF-QESA feature set on the ASVSpooof 2017 v2.0 dataset. Experiments are performed on additional datasets to show the generalizability of CFCCIF-QESA. The performance is compared with the existing feature sets in terms of EER, accuracy, and model-level information-theoretic measures, such as Kullback-Leibler Divergence (KLD) and Jensen-Shannon Divergence (JSD).

4.2.5.1 Initial Parameterization

4.2.5.2 Parameter Tuning on the ASVSpooof 2017 v2.0 Dataset

In this subsection, we present the results on the Dev set of the ASVSpooof 2017 v2.0 dataset to obtain the optimized parameters of the proposed feature set, such as the bandwidth of the subband filters used in the cochlear filterbank, the number of subband filters in the filterbank, and the dimension of the CFCCIF-QESA feature vector. Consequently, with the optimally tuned parameters, we experimentally determine the EER on the Eval set. Here, we will investigate the behaviour of EER on both Dev and Eval sets as the feature parameters are varied.

- Effect of Number of Subband Filters

The auditory system in humans comprises numerous subband filters, which form a dense filterbank [182]. To that effect, experiments were performed to observe the performance due to varying the number of subband filters in the cochlear filterbank. A speech signal filtered to a narrowband signal results in a more accurate estimation of IF. Given that CFCCIF-QESA is extracted from linearly-spaced subband filters, the number of subband filters explicitly affect the resolution in the frequency-domain. The number of subband filters should be large enough such that there is no loss of information. At the same time, considering a large number of subband filters results in complete overlap with adjacent subband filters.

Figure 4.8 (a) shows the effect of the number of subband filters on the EER. It should be noted that with 40 subband filters, an EER of 12.07% on the Dev set is obtained. Furthermore, the effect of increasing the number of subband filters is observed, and a minimum EER of 8.61% is achieved for 200 subband filters. This finding is in agreement with a recent study reported

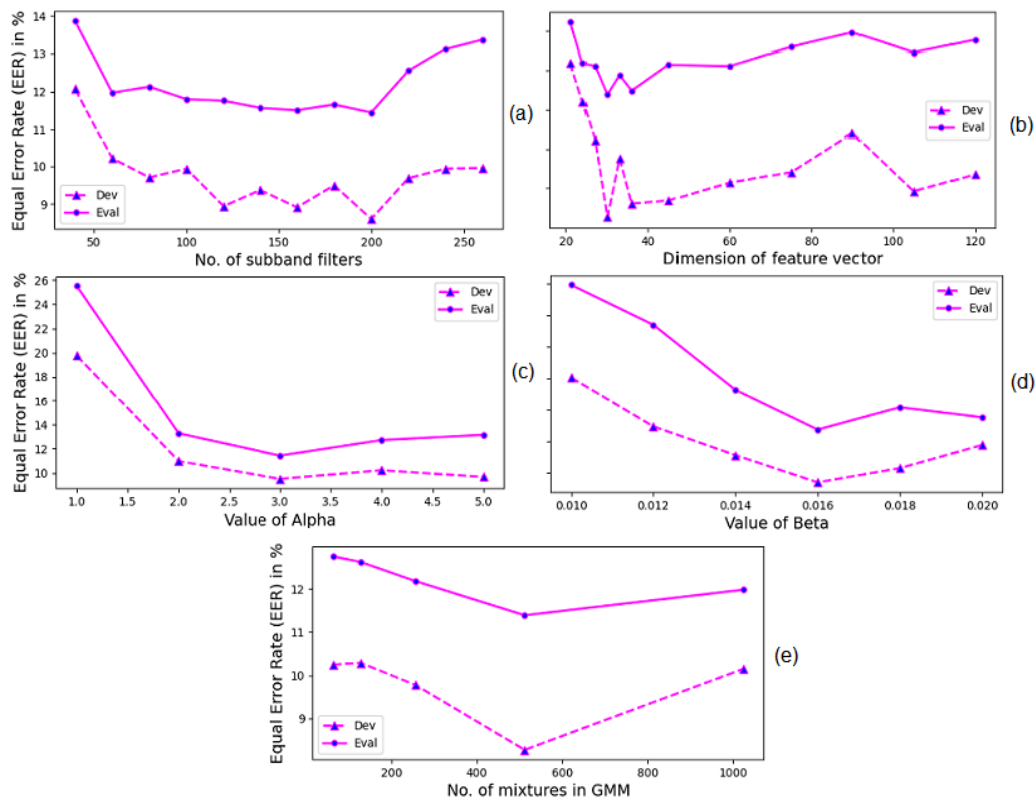


Figure 4.8: Results (in % EER) for the CFCCIF-QESA feature set on Dev and Eval set of ASVSpoof 2017 v2.0 dataset w.r.t. the (a) number of subband filters in the cochlear filterbank, (b) dimension of feature vector, (c) value of α , (d) value of β , and (e) number of mixtures in GMM.

in [183]. Overall, we are able to achieve 40% decrease in the EER by taking the number of subband filters as 200, as compared to 40 filters.

- Effect of Shape and Size Parameters of the Subband Filters

The *shape* and *width* of the cochlear filters in the filterbank is governed by α and β parameters of the subband filter (as shown in eq. (4.33)). The parameter α controls the center frequencies of the cochlear subband filters, and the parameter β controls the spread and ‘peakiness’ of the subband filters [80]. Figure 4.8 (c) and Figure 4.8 (d) show the effect of variation of α and β , respectively, on the EER. It was observed that for both the Dev and Eval sets, the minimum EER was achieved at $\alpha = 3$ (which is in agreement with the recent study reported in [70]) and $\beta = 0.016$. Additionally, in the original study reported in [80], it was found that the value of β can be selected depending on the application. In particular, the value of β should preferably be taken as 0.2 for the reduction in noise and even smaller for tasks such as feature extraction for pattern recognition [80]. The optimal value of

$\beta = 0.016$ for CFCCIF-QESA is in agreement with the study reported in [80]. To that effect, the values of subband filter parameters α and β are kept fixed at 3 and 0.016, respectively, to determine the effect of the dimension of the feature vector, as described next.

- Effect of Dimension of Feature Vector

The dimension (D) of the CFCCIF-QESA feature vector consists of static, Δ (delta), and $\Delta\Delta$ (delta-delta) coefficients. We performed experiments to estimate the optimal value of D . Figure 4.8 (b) shows the variation of % EER w.r.t. D varied from 18 to 120. On the Dev set, we obtain a minimum EER of 8.82% for $D = 30$ (i.e., 10 static coefficients, 10 Δ , and 10 $\Delta\Delta$ coefficients). Similarly, on the Eval set, we obtain a minimum EER of 11.57% for $D = 30$. The lower value of D also contributes to less computational time and resources. Theoretically, a higher value of D should allow more information to be captured. However, increasing the dimension of the feature vector can also increase the redundancy and noise [184]. In our experiments, we observed a decrease in performance with an increase in D . This behaviour is due to the case that as the dimension of the feature vector is increased, one introduces redundancy and features that are irrelevant to the class label (i.e., genuine and spoof class labels). This, in turn, degrades the performance of the classifier [184] implying that the classification error increases with the increase in the number of features. Adding to the problem, higher-dimensional feature vectors have exponentially increasing computational time. To that effect, our obtained results on lower dimension of 30 are well suited for the replay SSD task.

- Effect of Number of Mixtures in GMM

In order to select the optimal number of mixtures in GMM, we performed experiments by varying the number of mixtures in GMM using 30-D CFCCIF-QESA. Figure 4.8 (e) shows the effect of the GMM mixtures on the EER. We observe that for both Dev and Eval sets, the EER follows a similar trend. In particular, relatively minimum EER is achieved for 512 mixtures, for *both* Dev and Eval sets. The optimized EER thus achieved is 9.48% and 11.40% on Dev and Eval datasets, respectively, after fine-tuning all the parameters during the extraction of CFCCIF-QESA.

4.2.5.3 Results on the ASVSpooF 2017 v2.0 Database w.r.t. Various Classifiers

Given the fine-tuning of parameters of CFCCIF-QESA as discussed in subsection 4.2.5.2, the optimal CFCCIF-QESA feature set on the ASVSpooF 2017 v2.0 is 30-D, and gives the best results with 512 mixtures in GMM. Keeping these optimal parameters, the results of CFCCIF-QESA are compared with the other existing feature sets using GMM, CNN, LCNN, and ResNet as shown in Table 4.1, Table 4.2, Table 4.3, and Table 4.4, respectively.

Table 4.1: Results on the ASVSpooF 2017 v2.0 Database using GMM.

Feature Set	Dev		Eval	
	% EER	% Accu.	% EER	% Accu.
CQCC (S1)	12.87	81.75	18.81	59.72
CFCC (S2)	17.60	79.29	18.97	59.96
CFCCIF (S3)	16.61	78.59	17.38	58.13
CFCCIF-ESA (S4)	11.54	82.57	14.77	68.52
CFCCIF-QESA (S5)	9.48	87.30	11.40	73.35
S1+S5	9.48	87.30	11.40	73.35
S2+S5	9.47	87.34	11.39	73.36
S3+S5	9.37	87.34	11.38	73.39
S4+S5	9.25	87.70	11.31	73.80
S2+S3+S4+S5	9.22	87.90	11.25	73.96
S1+S2+S3+S4+S5	9.21	87.94	11.24	74.03
+ indicates score-level fusion as per eq. (3.7)				

It can be observed that our proposed feature set (denoted by S5) performs relatively the best as compared to the other systems (i.e., S1 to S4). To be specific, we achieve EER of 11.40% and accuracy of 73.35% on the ASVSpooF 2017 Eval set. To emphasize the benefit of incorporating the quadrature phase component, the results show that the proposed system S5 (i.e., with quadrature phase component) gives an absolute decrease in EER of 3.37% and an absolute improvement of 4.83% in accuracy, as compared to system S4 (with no quadrature phase component). Furthermore, we performed score-level fusion as per eq. (3.7) (denoted by + in Table 4.1) of system S5 with all the remaining systems S1 to S4. The score-level fusion of three systems, which are based on cochlear filtering (i.e., S3, S4, and S5) further reduced the EER to 9.36% and 11.19% on the Dev and Eval sets, respectively.

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

Table 4.2: Results on the ASVSpooof 2017 v2.0 Database using CNN.

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

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

Table 4.3: Results on the ASVSpooof 2017 v2.0 Database using LCNN.

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

Table 4.3 shows the performance when LCNN was used as the classifier. A relatively better performance of S5 with LCNN as compared to CNN is observed. In particular, we obtain an EER of 17.52% and an accuracy of 59.30% on the Eval set of ASVSpooof 2017 database. Furthermore, performance behaviour similar to GMM and CNN can be observed as S5 performs better, when compared to all the cochlear filter-based features (i.e., S2, S3, and S4). This also confirms the significance of quadrature phase component in the proposed feature set. Furthermore, if we compare the performance of individual feature sets (from S1 to S4), with their

individual fusion performance with S5 (i.e., S1+S5, S2+S5, S3+S5, and S4+S5), we observe improvement in the performance for *each* fusion case. To that effect, on the Eval set of ASVSpooF 2017, the maximum absolute decrease in EER of 15.11% and 22.89% in accuracy is observed w.r.t. S1 and S1+S5, as shown in the Table 4.3.

Table 4.4: Results on the ASVSpooF 2017 v2.0 Database using ResNet.

Feature Set	Dev		Eval	
	% EER	% Accu.	% EER	% Accu.
CQCC (S1)	6.05	90.52	21.72	46.64
CFCC (S2)	4.21	95.78	23.34	57.71
CFCCIF (S3)	15.00	84.50	21.87	69.99
CFCCIF-ESA (S4)	11.49	87.07	22.40	70.07
CFCCIF-QESA (S5)	11.57	87.54	16.71	73.52
S1+S5	4.09	91.16	14.90	74.34
S2+S5	3.04	96.54	13.69	75.69
S3+S5	11.07	88.33	16.68	73.63
S4+S5	9.97	88.47	16.68	73.62
S2+S3+S4+S5	2.95	97.07	13.04	75.96
S1+S2+S3+S4+S5	2.33	97.54	12.88	76.35

Table 4.4 shows the performance of the system, when ResNet was used as the classifier. A relatively better performance of S5 is observed with ResNet as compared to the LCNN and CNN. In particular, we obtain an EER of 16.71% and an accuracy of 73.52% on the Eval set of ASVSpooF 2017 database. It should be noted that S5 outperforms all the other feature sets, including the state-of-the-art CQCC feature set.

Classifier-Level Fusion: Given various classifiers (i.e., GMM, CNN, LCNN, and ResNet) were used on the ASVspooF 2017 v2.0 dataset, we now present the classifier-level fusion results in Table 4.5. It shows the results obtained on the proposed CFCCIF-QESA using GMM, CNN, LCCN, and ResNet labelled as S1, S2, S3, and S4, respectively. The best performance on the Eval set is observed when the scores of all the four classifiers are fused, leading to an EER of 10.99%. Notably, CFCCIF-QESA shows relatively the best performance using GMM. The better performance of GMM can be due to the data being more approximated to be Gaussian and the acoustic characteristics are better suited for GMM.

4.2.5.4 Analysis of Latency on the ASVSpooF 2017 v2.0 Database

Latency period represents the performance evaluation in terms of %EER *w.r.t* different durations of speech segment in an utterance. The utterance duration ranges

Table 4.5: Results of Classifier-Level Fusion of the CFCCIF-QESA Feature Set using Different Classifiers on the ASVSpooF 2017 v2.0 Dataset.

Classifier Used	Dev		Eval	
	% EER	% Accu.	% EER	% Accu.
GMM (S1)	9.48	87.30	11.40	73.35
CNN (S2)	9.74	88.36	19.10	57.40
LCNN (S3)	11.22	87.10	17.52	59.30
ResNet (S4)	11.57	87.54	16.71	73.52
S1+S2	6.97	90.00	11.40	73.35
S1+S3	7.79	89.82	11.00	74.00
S2+S3	8.78	89.75	16.55	66.36
S3+S4	9.26	88.45	16.10	66.65
S1+S2+S3	6.62	90.99	11.00	74.02
S1+S2+S3+S4	6.62	90.99	10.99	74.10

from 20 ms to 2 seconds, with an interval of 200 ms. Further, the utterance duration is selected by considering the number of frames. Figure 4.9 shows the comparison between the CQCC baseline, CFCCIF, CFCCIF-ESA, and CFCCIF-QESA. It can be observed that all the feature sets show comparable latency with each other for the Dev set of ASVSpooF 2017 as shown in Figure 4.9. However, for the Eval set of ASVSpooF 2017 as shown in Figure 4.9 (b), we observe a considerable improvement of CFCCIF, CFCCIF-ESA, and CFCCIF-QESA in latency performance w.r.t. the CQCC baseline. Furthermore, the %EER converges to the minimum value as the speech duration provided to the model of the SSD system increases. Additionally, the feature performance is better if for a low latency period, the %EER is also low, indicating faster classification by the model and thus, indicating the suitability of the system for practical deployment.

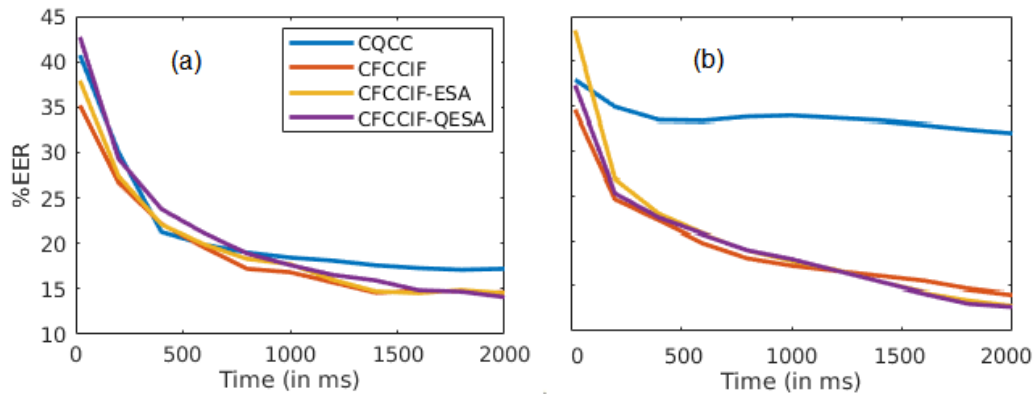


Figure 4.9: Analysis of latency period for the SSD system (a) Dev set, and (b) Eval set of ASVSpooF 2017 dataset using various feature sets.

4.2.5.5 Cross-Database Evaluation with Training on ASVSpooF 17 V2.0

For the majority of existing SSD systems, training and evaluation are done on the same database. Such type of evaluation procedures are known as *self-classification* [185]. However, self-classification does not represent true generalization capabilities of the SSD system in realistic scenarios of ASV. Historically, the original study on CFCC reported its near performance of CFCC with MFCC under matched condition. However, the performance of CFCC was found to be significantly improved compared to its MFCC counterpart under mismatched conditions of training (clear) *vs.* testing (noisy) [81]. Therefore, in order to perform cross-database Eval, using GMM as the classifier, we train using ASVSpooF 2017 v2.0 training set, and test on the Dev and Eval sets of different datasets. In this work, we use ASVSpooF 2019 PA (Dev and Eval sets), BTAS 2016 (Dev and Eval sets), and the VSDC (Eval set). In this subsection, we will discuss the experimental results on cross-database evaluation using GMM as classifier for the mentioned datasets.

- **Training on the ASVSpooF 2017 v2.0, and testing on the ASVSpooF 2019 PA Dev and Eval sets:** Table 4.6 shows the experimental results on cross-dataset evaluation between ASVSpooF 2017 v2.0 and ASVSpooF 2019 PA datasets. It is worth noting that even though both the datasets correspond to the same attack (i.e., replay), the performance of the countermeasure system is not that remarkable. In particular, the state-of-the-art features, such as CQCC, CFCC, and CFCCIF show comparable performance of 50% EER. Furthermore, the CFCCIF-ESA and CFCCIF-QESA feature sets show relatively better performance (but not at all good) with EER on the Eval set as 42.64% and 42.85%, respectively, however, still there is a scope for further improvement. The significant degradation in the performance is primarily due to a mismatch in the acoustics of real and simulated replay spoofed signals for training and testing (or vice-versa).

Table 4.6: Results on Cross-Database Evaluation Between ASVSpooF 2017 v2.0 and ASVSpooF 2019 PA Corpora.

Training Dataset	ASVSpooF 2017 v2.0			
	ASVSpooF 2019 PA			
	Dev		Eval	
Feature Set	%EER	%Accu.	%EER	%Accu.
CQCC	48.50	18.21	49.99	13.42
CFCC	49.94	18.18	49.99	13.42
CFCCIF	49.99	20.43	50	17.47
CFCCIF-ESA	47.07	33.06	42.64	36.57
CFCCIF-QESA	44.48	35.13	42.85	32.72

- **Training on the ASVSpooof 2017 v2.0, and testing on BTAS Dev and Eval Sets:** Table 4.7 shows the experimental results on cross-dataset evaluation between ASVSpooof 2017 v2.0 and BTAS 2016 datasets. It can be observed that IF-based features perform *significantly* better as compared to the CQCC and CFCC feature sets. This signifies the importance of IF in replay attack detection, even under mismatched conditions.

Table 4.7: Results on Cross-Database Evaluation Between ASVSpooof 2017 v2.0 and BTAS 2016 Corpora.

Training Dataset	ASVSpooof 2017 v2.0			
Testing Dataset	BTAS 2016			
	Dev		Eval	
Feature Set	%EER	%Accu.	%EER	%Accu.
CQCC	49.40	14.59	48.04	19.20
CFCC	49.57	11.46	49.10	11.46
CFCCIF	12.55	90.45	15.51	86.56
CFCCIF-ESA	11.60	92.09	14.45	89.05
CFCCIF-QESA	13.05	91.03	15.51	87.40

- **Training on the ASVSpooof 2017 v2.0, and testing on VSDC dataset:** The ASVSpooof 2017 v2.0 dataset consists of 1st order replay recordings. On the other hand, the VSDC dataset consists of the recordings of 1st order and 2nd order replay scenarios. A majority of the existing SSD systems are designed using self-classification. However, true generalization capabilities of the SSD system cannot be evaluated by self-classification. Historically, the original study on CFCC reported its near performance of CFCC with MFCC under matched condition. However, the performance of CFCC was found to be significantly improved compared to its MFCC counterpart under mismatched conditions of training (clear) *vs.* testing (noisy) [81].

Table 4.8: Results on Cross-Database Evaluation Between ASVSpooof 2017 v2.0 Corpus and VSDC Corpus.

Training Dataset →	ASVSpooof 2017 v2.0 (Train)				VSDC (Full Data)			
Testing Dataset →	VSDC (Full Data)				ASVSpooof 2017 v2.0 (Dev & Eval)			
	OPR-1PR		OPR-2PR		Dev		Eval	
Feature Set ↓	% EER	% Accu.	% EER	% Accu.	% EER	% Accu.	% EER	% Accu.
CQCC	52.12	48.33	43.16	55.68	32.52	55.20	35.30	51.23
CFCC	47.48	51.10	48.89	44.52	29.30	62.02	32.12	54.89
CFCCIF	44.23	53.25	35.43	63.35	28.34	66.19	31.54	57.26
CFCCIF-ESA	43.42	55.32	32.00	67.65	25.82	74.56	29.46	60.80
CFCCIF-QESA	42.63	56.54	31.40	68.78	25.40	75.20	26.56	65.30

Therefore, in order to perform cross-database evaluation, using GMM as the classifier, we perform two sets of experiments (as shown in Table 4.8) - one by

training using the ASVSpooF 2017 v2.0 training dataset, the second by training using the complete VSDC dataset. When training on the ASVSpooF 2017 v2.0 training dataset, we performed evaluation on two cases: genuine *vs.* 1st order replay (i.e., 0PR *vs.* 1PR), and the other on genuine *vs.* 2nd order replay (i.e., 0PR *vs.* 2PR). It can be observed from Table 4.8 that even under mismatched condition of training and testing, our proposed feature set outperforms the rest of the feature sets. In particular, as compared to the CQCC, an absolute decrease in EER of 9.49% and 11.76% is observed for 0PR-1PR and 0PR-2PR settings, respectively, due to the proposed CFCCIF-QESA. For the second set of experiments, where training was done on VSDC dataset and evaluation was done on the ASVSpooF 2017 v2.0 dataset, an absolute decrease in EER of 8.74% was observed by CFCCIF-QESA as compared to the CQCC feature set.

4.2.5.6 SSD System Performance Under Ideal Conditions

To evaluate the performance under ideal scenarios, we performed experiments on the ASVSpooF 2017 v2.0 dataset for 2 scenarios: case 1) when the system is not under attack, and case 2) when it is under attack. For case-1, when the system is not under attack, it means that the inputs to the system are strictly genuine signals. To do so, we evaluated the system’s performance by considering only genuine speech as shown in Table 4.9. An ideal system will accept *all* the genuine utterances, and hence, the False Rejection Rate (FRR) will be 0. This means that no genuine utterance will be falsely classified as spooF, in an ideal system. For case-2,

Table 4.9: System Performance When it is Not Under Attack.

Feature Set	Accuracy in %	
	Dev	Eval
CQCC	95.00	92.52
CFCC	94.21	92.44
CFCCIF	96.18	93.37
CFCCIF-ESA	97.10	93.14
CFCCIF-QESA	93.55	93.22

when the system is under attack, it means that the inputs to the ASV system are strictly spooFed signals. To do so, we evaluated the system’s performance only on spooFed speech, as shown in Table 4.10. An ideal system will reject *all* the spooFed utterances, and hence the False Acceptance Rate (FAR) will be 0. This means that no spooFed utterance will be falsely classified as genuine in an ideal system.

Table 4.10: System Performance When it is Under Attack.

Feature Set	Accuracy in %	
	Dev	Eval
CQCC	71.15	56.17
CFCC	67.36	56.45
CFCCIF	64.52	54.32
CFCCIF-ESA	70.94	65.86
CFCCIF-QESA	84.21	71.20

4.2.5.7 Results on the ASVSpooF 2019 PA Database

The ASVSpooF 2019 PA dataset contains a controlled simulation of replay attacks on which we performed experiments, whose results are presented in Table 4.11. It can be observed from Table 4.11 that the CFCCIF-QESA feature set achieves

Table 4.11: Results on the ASVSpooF 2019 PA Database using GMM.

Feature Set	Dev		Eval	
	% EER	% Accu.	% EER	% Accu.
CQCC (S1)	9.87	90.75	11.04	87.72
CFCC (S2)	17.60	85.29	18.97	82.96
CFCCIF (S3)	36.93	66.13	37.61	60.13
CFCCIF-ESA (S4)	36.29	65.57	36.94	61.23
CFCCIF-QESA (S5)	22.40	80.05	25.71	75.35
S1+S5	9.85	90.89	11.01	87.75
S2+S5	17.55	85.50	18.75	83.10
S3+S5	22.20	80.20	25.60	75.50
S4+S5	22.18	80.22	25.58	75.60
S2+S3+S4+S5	17.40	85.70	18.55	83.60
S1+S2+S3+S4+S5	9.43	91.03	10.89	88.02

an EER of 25.71% and an accuracy of 75.35% on the Eval set. Even though the performance of CQCC remains to be relatively better, it should be noted that the CFCCIF-QESA performs better than the CFCCIF-ESA. To that effect, on the Eval set, we achieve an absolute decrease of 11.23% in EER, and an improvement of 14.12% in accuracy is achieved.

4.2.5.8 Results on the BTAS 2016 Dataset

The BTAS 2016 dataset is an extended version of the ASVSpooF 2015 dataset. In particular, it contains VC, SS, and replay spoofed utterances. However, the other three datasets (i.e., ASVSpooF 2017 v2.0, ASVSpooF 2019, and VSDC) used in this study contains only the replay spoofed speech. Hence, we have again fine-tuned

the cochlear filter parameters α and β on the BTAS 2016 Dataset. It should be observed that the values of the optimized filter parameters α and β are also data-dependent. The fine-tuned results are shown in the Fig. 4.10, giving us the optimal value of $\alpha = 5$ and $\beta = 0.014$. Based on these fine-tuned parameters, ex-

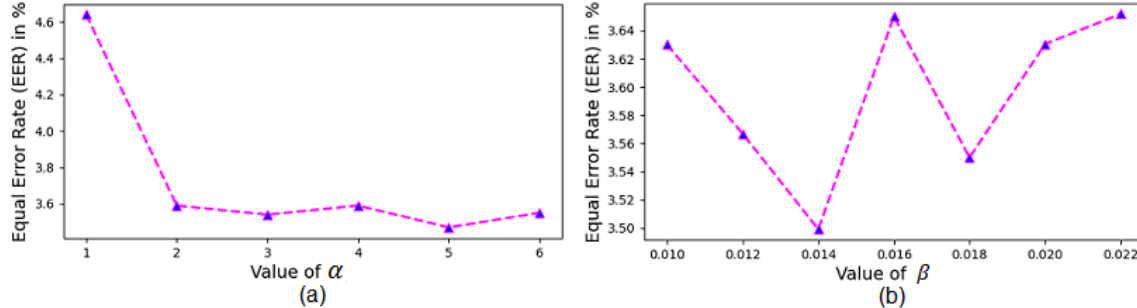


Figure 4.10: Results (in % EER) on Dev set of BTAS 2016 dataset with variation of (a) value of α , and (b) value of β .

periments are performed and the results are shown in Table 4.12. It can be ob-

Table 4.12: Results (in % EER and Accuracy) on the BTAS 2016 Dataset using GMM.

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

served that CFCCIF-QESA performs relatively close to the CFCCIF-ESA feature set. However, we observe the best performance in EER, when all the features are fused to give an EER of 3.43% and an accuracy of 93.67%.

4.2.5.9 Results on the ReMASC Dataset

The ReMASC dataset was developed for anti-spoofing of VAs. The CFCCIF-QESA feature set is evaluated on the ReMASC dataset. Table 4.13 shows the

environment-wise performance in terms of % EER. The analysis shows the enhanced SSD capability of the CFCCIF-QESA feature set over CFCCIF-ESA, particularly for the case of Env B, Env C, and Env D. Furthermore, for Env A, the performance of CFCCIF-ESA and CFCCIF-QESA is comparable. Next, Table 4.14

Table 4.13: Environment-wise Results (in % EER) Using GMM as the Classifier.

Feature Set	Environment-wise % EER			
	Env A	Env B	Env C	Env D
CQCC	22.58	48.48	43.45	14.47
CFCCIF-ESA	29.15	46.33	47.02	15.13
CFCCIF-QESA	29.22	45.53	46.81	12.75
CQCC \oplus CFCCIF-ESA	22.56	46.33	43.45	12.42
CQCC \oplus CFCCIF-QESA	22.58	45.53	43.45	11.55
\oplus indicates score-level fusion as per eq. (3.7)				

shows overall results in terms of %EER on GMM and CNN classifiers. It can be observed that in the case of classification by GMM, CFCCIF-QESA achieves an overall EER of 28.71% on the Eval set, thereby achieving relative % decrease of 4.17% w.r.t. CFCCIF-ESA.

Table 4.14: Results in %EER using GMM and CNN as the Classifiers.

Feature Set	GMM		CNN	
	Dev	Eval	Dev	Eval
CQCC	19.94	22.56	15.36	25.33
CFCCIF-ESA	21.64	29.95	16.23	28.03
CFCCIF-QESA	26.93	28.71	15.47	29.89
CQCC \oplus CFCCIF-ESA	18.14	22.56	11.20	23.55
CQCC \oplus CFCCIF-QESA	18.69	22.02	10.63	23.84

Figure 4.11 shows the Detection Error Trade-off (DET) curve obtained on the Dev, and Eval sets of the ReMASC dataset. Furthermore, the latency period for CFCCIF-QESA w.r.t the CFCCIF-ESA feature set is also investigated. Such analysis enables us to investigate how fast the SSD system is w.r.t. deployment in real-world applications. Latency is the performance evaluation in terms of %EER w.r.t different durations of speech segment in an utterance. Figure 4.12 shows the latency analysis for the CFCCIF-ESA and the CFCCIF-QESA feature sets. The utterance duration ranges from 10 ms to 60 ms, with an interval of 10 ms. From Figure 4.12, it can be observed that the proposed CFCCIF-QESA shows relatively smaller latency as compared to CFCCIF-ESA, i.e., the CFCCIF-QESA feature set

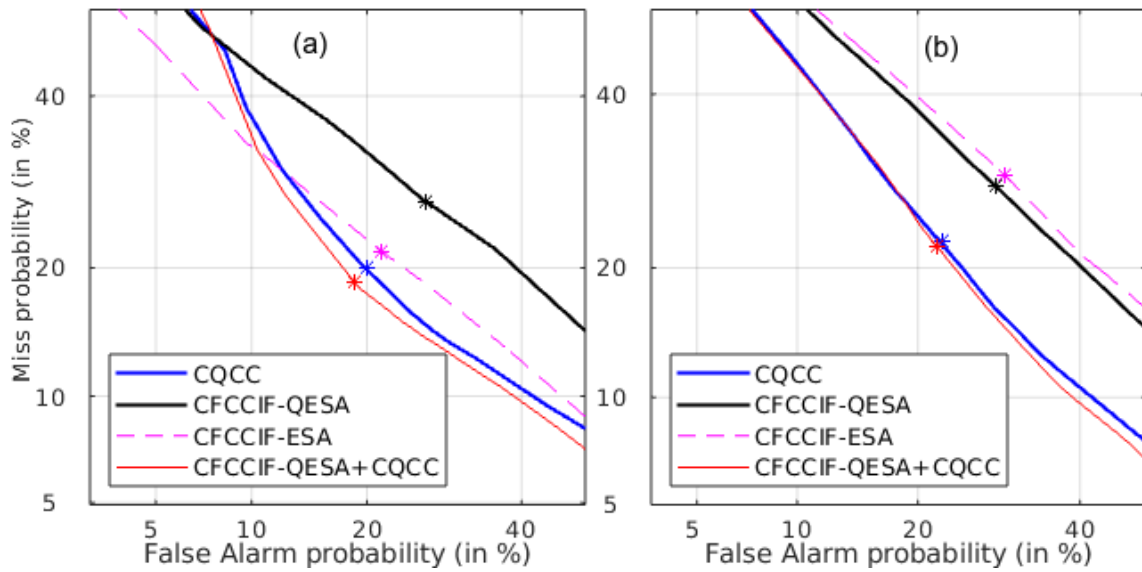


Figure 4.11: DET curves on (a) Dev set, and (b) Eval set of ReMASC dataset.

reaches better performance than the CFCCIF-ESA feature set even for a short duration (i.e., 10 ms) of speech. The low latency achieved by CFCCIF-QESA indicates the ability of faster classification by the model and thus, better suitability of CFCCIF-QESA for practical system deployment as compared to CFCCIF-ESA.

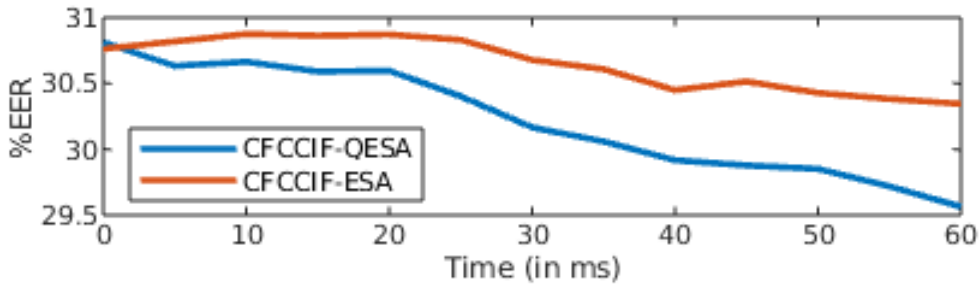


Figure 4.12: Latency curves for CFCCIF-ESA vs. CFCCIF-QESA on ReMASC dataset.

4.2.5.10 Results on the ASVSpooF 2015 dataset

Table 4.15 shows the results w.r.t. the 10 different types of SS and VC attacks, labelled from S1 to S10. Out of these S1 to S5 are *known* attacks, and S6 to S10 are *unknown* attacks. The average EER is denoted as AEER. It can be observed that in the case of known attacks, CFCCIF-QESA achieves relatively better AEER of 0.31%, than CFCCIF-ESA. The achieved % relative decrease in EER is 27.90%.

Table 4.15: Results (in % EER and % AEER) on the ASVSpooof 2015 dataset for various feature sets using GMM as a classifier.

Feature Set	Known Attacks					Unknown Attacks					All		
	S1	S2	S3	S4	S5	AEER	S6	S7	S8	S9	S10	AEER	AEER
LFCC-DA [186]	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 [187]	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 [70]	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 [70]	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 [70]	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 [188]	0.11	1.23	0.03	0.05	0.75	0.43	0.80	0.31	0.89	0.63	2.40	1.00	0.71
CFCCIF-QESA	0.09	0.74	0.03	0.05	0.67	0.31	0.38	0.14	1.34	0.40	1.70	0.79	0.55

Furthermore, in the case of unknown attacks, CFCCIF-QESA achieves AEER of 0.79%, thereby leading to a reduction of 21% in AEER w.r.t. the CFCCIF-ESA feature set. In particular, for the case of S10 attack, CFCCIF-QESA achieves an EER of 1.70%, thereby leading to relative decrease of 29.16% in EER w.r.t. CFCCIF-ESA. Moreover, it should also be noted that amongst the cochlear filterbank-based feature sets, CFCCIF-QESA gives the best performance of 0.55% AEER of all the attacks as shown in Table 4.15. In addition, CFCCIF-QESA also shows better generalization ability in the classification of known as well as unknown attacks.

Furthermore, we show the DET curve to observe the performance across all the operating points of the SSD system [189]. Figure 4.13 shows DET curves of CFCCIF-ESA, CFCCIF-QESA, and their score-level fusion. For Dev set, $\alpha = 0.68$, and for Eval set $\alpha = 0.64$. It can be observed that the fusion performs relatively the best at all the operating points throughout the curve and hence, this dictates that complementary information between CFCCIF-ESA and CFCCIF-QESA is captured effectively.

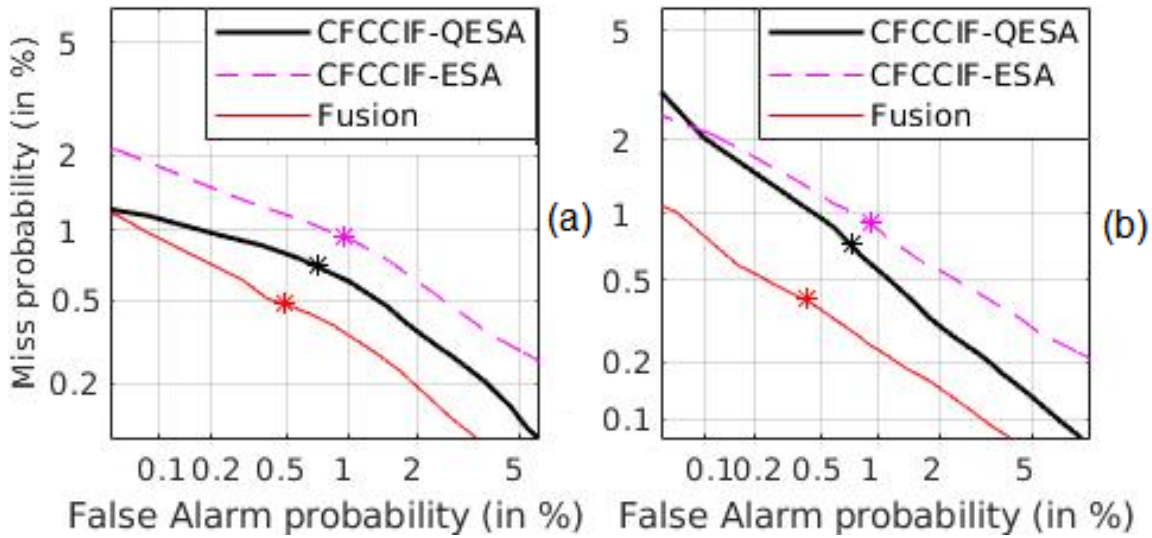


Figure 4.13: DET curves (a) Dev, and (b) Eval set.

We also investigate the latency period for CFCCIF-QESA *w.r.t* the other feature sets. Such an analysis enables us to investigate how fast the SSD system is *w.r.t* deployment in real-world applications. Latency is the performance evaluation in terms of %EER *w.r.t* different durations of speech segment in an utterance. To that effect, experiments were performed on INTEL(R) Core(TM) i5-2400 CPU at 3.10 GHz. For estimation of latency, the scores of all the utterances in the Dev and Eval sets of the ASVspoof 2015 dataset were used. Figure 4.14 (a) and (b) show the latency analysis on the Dev and Eval sets, respectively. The utterance duration ranges from 20 ms to 2 seconds, with a step-size of 500 ms. From Figure 4.14

(a), it can be observed that the proposed CFCCIF-QESA shows relatively smaller latency as compared to CFCCIF-ESA, i.e., CFCCIF-QESA reaches better performance than the CFCCIF-ESA feature set even for a short duration (i.e., 200 ms) of speech. Furthermore, for the case of the Eval set, as shown in Figure 4.14 (b), both the features perform comparably to each other. The low latency achieved by CFCCIF-QESA on the Dev set indicates the ability of faster classification by the model and thus, better suitability of CFCCIF-QESA for practical SSD system deployment as compared to the CFCCIF-ESA feature set.

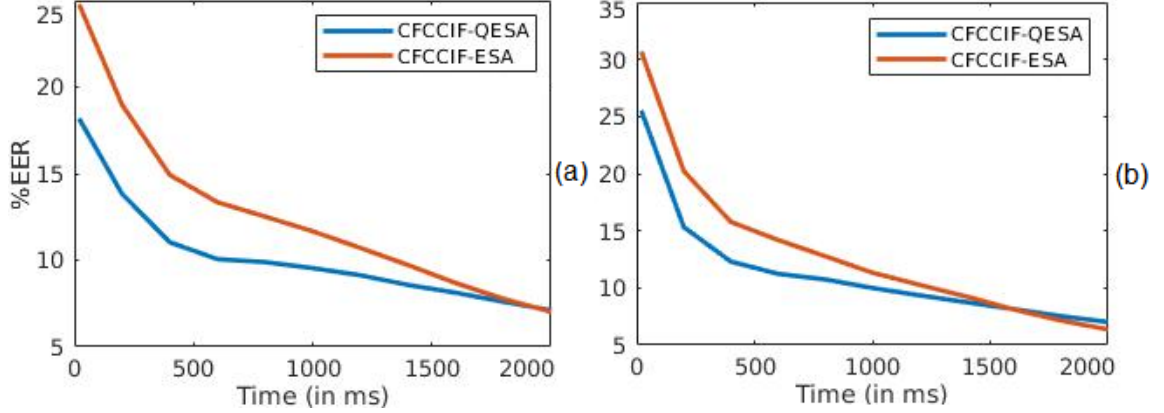


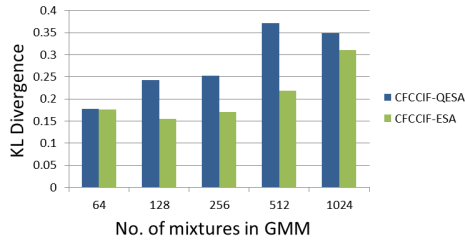
Figure 4.14: Latency curves (a) Dev, and (b) Eval set of the ASVSpooof 2015 dataset.

4.2.5.11 Analysis Using Model-Level Measures

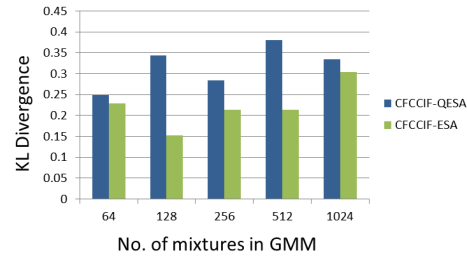
In addition to EER and accuracy, we present model-level information-theoretic measures to investigate the importance of the quadrature phase, in the proposed CFCCIF-QESA feature set. To that effect, in this subsection, we present two model-level measures, namely, Kullback-Leibler Divergence (KLD) and Jensen-Shannon Divergence (JSD) for the CFCCIF-ESA *vs.* CFCCIF-QESA feature sets.

KLD is an information-theoretic measure, which tells us about how much one Probability Distribution Function (PDF) differs from the other. KLD has been used extensively for analysis of different PDFs and the difference between them, in particular, its recent application in generative adversarial networks (GAN) literature [48, 190, 191] and in anti-spoofing, for ASV [192]. For the SSD task, it has been used as a model-level measure for distinguishing between genuine and spoof class [193]. If p and q are two PDFs, then it is a measure of how much information is lost, when $q(x)$ is used to approximate $p(x)$. Mathematically, it is expressed as [194]:

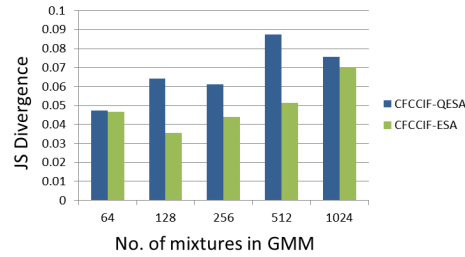
$$KLD(p||q) = - \int p(x) \ln q(x) dx - \left(- \int p(x) \ln p(x) dx \right), \quad (4.36)$$



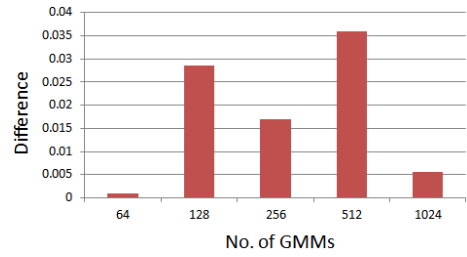
(a) KLD between genuine and spoof GMM



(b) KLD between spoof and genuine GMM

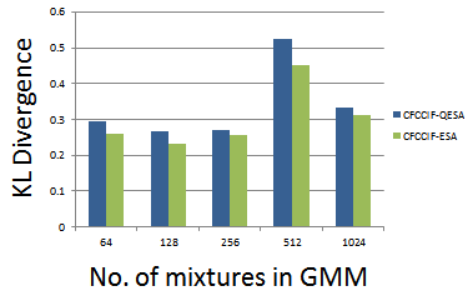


(c) JSD between genuine and spoof GMM

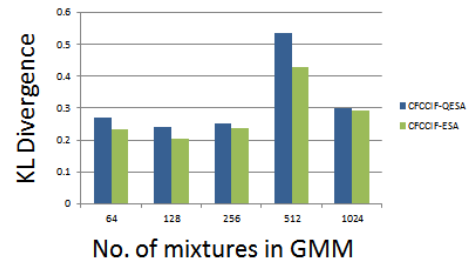


(d) Difference of JSD between CFCCIF-QESA and CFCCIF-ESA

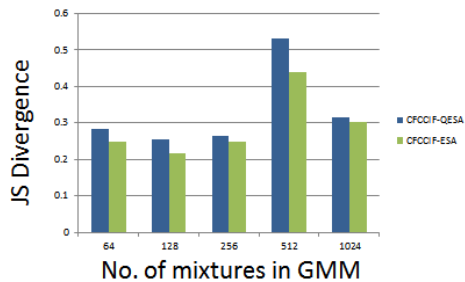
Figure 4.15: Comparative analysis of KLD and JSD for CFCCIF-ESA *vs.* CFCCIF-QESA for various numbers of mixtures used in GMM on the ASVspoof 2017 training corpus.



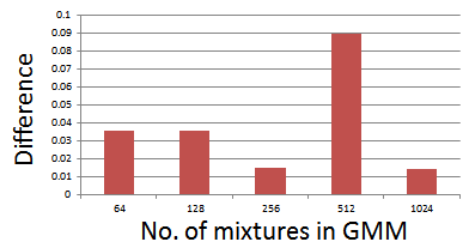
(a) KLD between genuine and spoof GMM



(b) KLD between spoof and genuine GMM



(c) JSD between genuine and spoof GMM

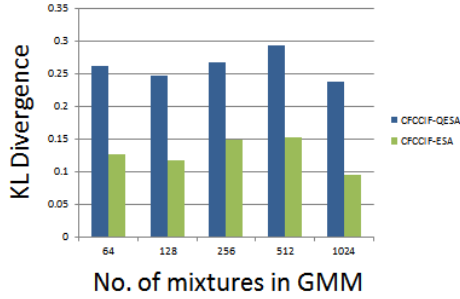


(d) Difference of JSD between CFCCIF-QESA and CFCCIF-ESA

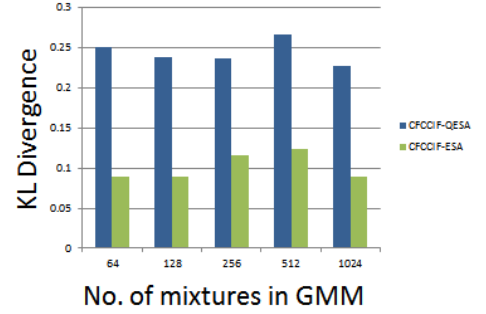
Figure 4.16: Comparative analysis of KLD and JSD for CFCCIF-ESA *vs.* CFCCIF-QESA for various numbers of mixtures used in GMM on the ASVspoof 2019 PA training corpus.

$$KLD(p||q) = - \int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx. \quad (4.37)$$

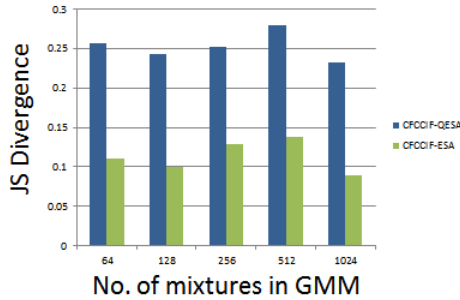
It is an *asymmetric* measure, i.e., $KLD(p||q) \neq KLD(q||p)$.



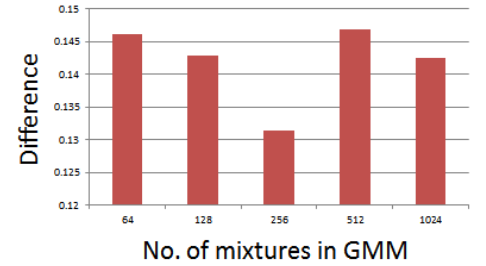
(a) KLD between genuine and spoof GMM



(b) KLD between spoof and genuine GMM



(c) JSD between genuine and spoof GMM



(d) Difference of JSD between CFCCIF-QESA and CFCCIF-ESA

Figure 4.17: Comparative analysis of KLD and JSD for CFCCIF-ESA *vs.* CFCCIF-QESA for various numbers of mixtures used in GMM on BTAS 2016 training corpus.

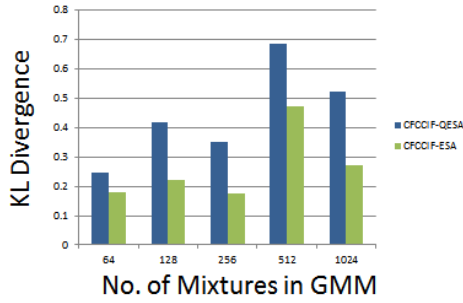
Furthermore, to eliminate the asymmetry between $KLD(p||q)$ and $KLD(q||p)$, we estimate the JSD. It is expressed as [194]:

$$JSD(p||q) = \frac{1}{2}KLD(p||m) + \frac{1}{2}KLD(q||m), \quad (4.38)$$

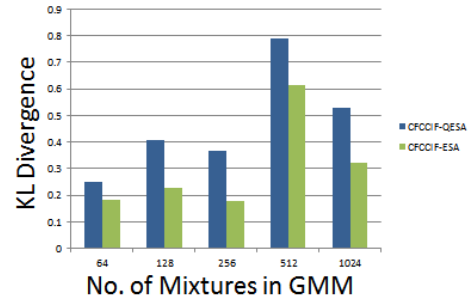
where m is estimated as $\frac{1}{2}(p + q)$. JSD is more useful as a measure, as it provides a smoothed and normalized version of KLD and hence, it is used in the original GAN literature as well [195].

In this work, the KLD and JSD between statistical GMM of genuine and spoofed speech, is used as a model-level measure having discriminative ability. To that effect, we have estimated KLD and JSD between genuine and spoof GMMs, P and Q . KLD is estimated between P and Q of the two GMMs corresponding to the genuine and the spoofed class, as shown in Algorithm 3.

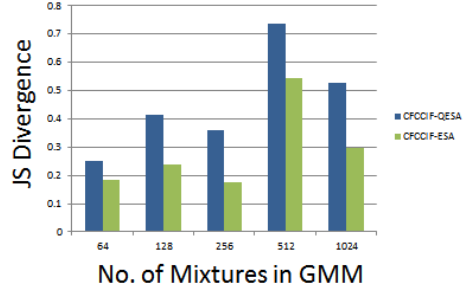
Here, we have experimentally emphasized the importance of the quadrature



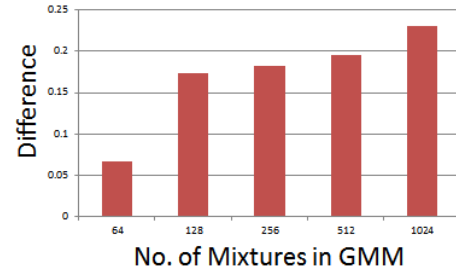
(a) KLD between genuine and spoof GMM



(b) KLD between spoof and genuine GMM



(c) JSD between genuine and spoof GMM



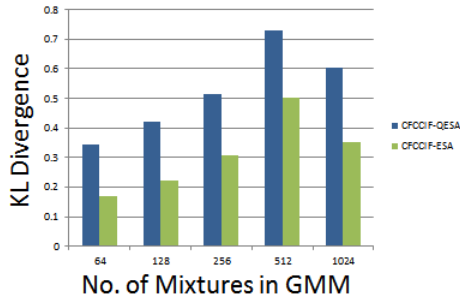
(d) Difference of JSD between CFCCIF-QESA and CFCCIF-ESA

Figure 4.18: Comparative analysis of KLD and JSD for CFCCIF-ESA *vs.* CFCCIF-QESA for various numbers of mixtures used in GMM on VSDC training corpus (only 0PR-1PR).

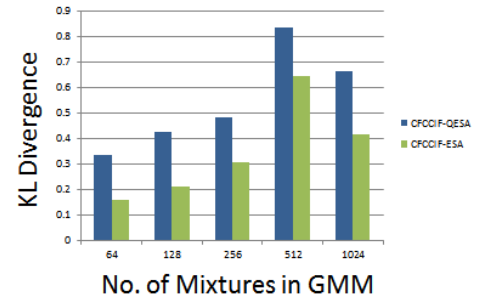
Algorithm 3 $KLD(P||Q)$

- 1: **procedure** $KLD((P||Q))$ $\triangleright P$ and Q are the GMMs of each class
 - 2: Check if the P and Q are valid probability distributions
 - 3: $KLD = \text{nansum}(P .* \log_2(P./Q))$
 - 4: **end procedure**
-

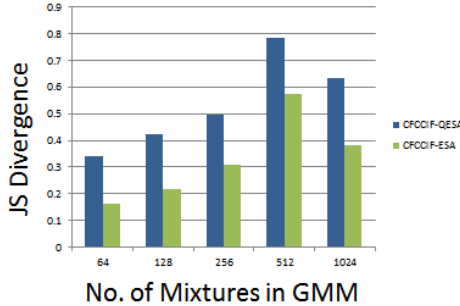
phase in CFCCIF-QESA for the SSD task. To that effect, we have estimated the KLD between the genuine and the spoofed GMMs of the two feature sets CFCCIF-ESA (i.e., without quadrature phase) and CFCCIF-QESA (i.e., with quadrature phase). Figure 4.15 (a) shows the KLD between two PDFs obtained from the training dataset of ASVSpooF 2017 v2.0 corpus for genuine and spoof GMMs, respectively. Given that KLD is an asymmetric measure, we take the converse case of Figure 4.15 (a). To that effect, Figure 4.15 (b) shows the KLD between spoof and genuine GMMs. Supporting our proposed argument of the importance of the quadrature phase, we observe that in *both* the cases (as shown in Figure 4.15(a) and Figure 4.15 (b)), the KLD between the GMMs of CFCCIF-QESA is *more* as compared to the KLD between the GMMs of CFCCIF-ESA and hence, confirming better discriminative ability of CFCCIF-QESA as compared to the CFCCIF-



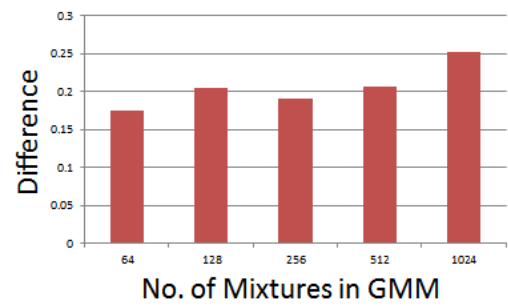
(a) KLD between genuine and spoof GMM



(b) KLD between spoof and genuine GMM



(c) JSD between genuine and spoof GMM

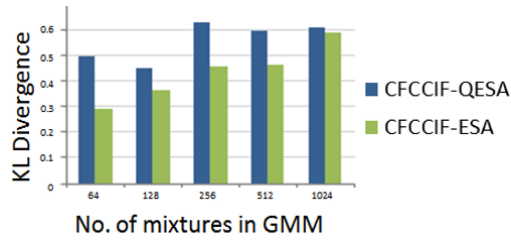


(d) Difference of JSD between CFCCIF-QESA and CFCCIF-ESA

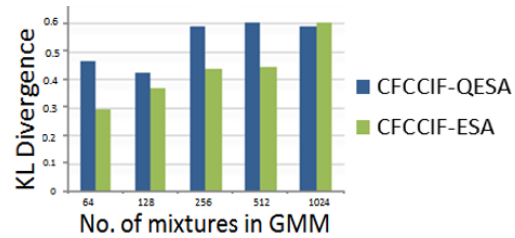
Figure 4.19: Comparative analysis of KLD and JSD for CFCCIF-ESA *vs.* CFCCIF-QESA for various numbers of mixtures used in GMM on VSDC training corpus (only 0PR-2PR).

ESA. Therefore, this analysis further reinforces the significance of our proposed inclusion of the quadrature phase in the CFCCIF-ESA framework to derive the CFCCIF-QESA feature set For the SSD task. Likewise, it can be observed from Figure 4.15 (c) that the JSD between genuine and spoof GMMs is higher for CFCCIF-QESA as compared to the CFCCIF-ESA. Furthermore, the difference in JSD, as shown in Figure 4.15 (d), is the highest for 512 mixtures used in GMM justifying relatively the best performance for GMM with 512 mixtures for the replay SSD tasks on the ASVSpooF 2017 v2.0 corpus (as also shown by the analysis from Figure 4.8(e) in subsection 4.2.5.2).

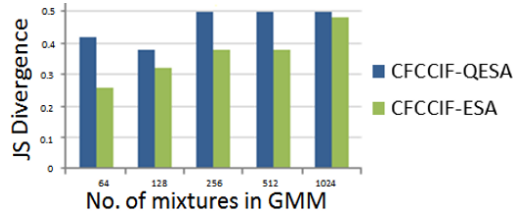
Similar analysis is performed on the training dataset of ASVSpooF 2019 PA, BTAS 2016, VSDC (1PR scenario), VSDC (2PR scenario), and ReMaSC, as shown in Figures 4.16, 4.17, 4.18, 4.19, 4.20, respectively . Interestingly, similar behaviour of KLD and JSD is observed across *all* these datasets, wherein, the CFCCIF-QESA feature set shows more discriminative ability as compared to CFCCIF-ESA, indicating generalizability of the proposed CFCCIF-QESA approach over the other datasets. For training sets of ASVSpooF 2017 v2.0, ASVSpooF 2019 PA, and BTAS, it can be observed that the best discriminative ability is achieved for 512 mixtures



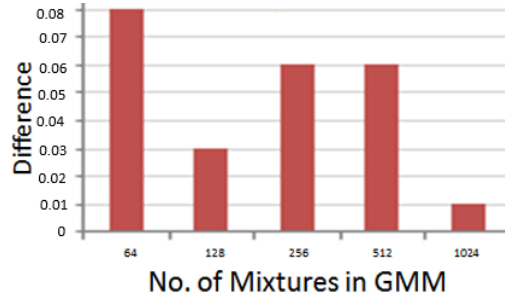
(a) KLD between genuine and spoof GMM



(b) KLD between spoof and genuine GMM



(c) JSD between genuine and spoof GMM



(d) Difference of JSD between CFCCIF-QESA and CFCCIF-ESA

Figure 4.20: Comparative analysis of KLD and JSD for CFCCIF-ESA *vs.* CFCCIF-QESA for various numbers of mixtures used in GMM on ReMASC training corpus.

in GMM. For the case of VSDC dataset, it is achieved for 1024 mixtures, while for ReMASC dataset, it is achieved for 64 mixtures in the GMM.

In subsection 4.2.3.1, an MI-based analysis was shown w.r.t. importance of quadrature relative phase as opposed to non-quadrature values of the relative phase (as shown in Figure 4.2). Consequently, we can also relate KLD with MI [177]. In particular,

$$I(X;Y) = KLD(p(x,y)||p(x)p(y)). \quad (4.39)$$

It is worth noting that the results for model-level KLD and JSD are in agreement with MI and classification-level EER results presented in this work.

4.3 Optimized Linear Frequency Residual Cepstral Coefficients (LFRCC)

Linear Prediction (LP) of speech has been widely used in many applications, from speech coding to analyzing excitation source-based information. The excitation source-based information is also known to carry speaker-specific information [15,

117,196–198]. The frequency response characteristics of the microphone, replay device, and acoustic environment are bandpass in nature. Due to the bandpass nature, the spectrum of the LP residual of replay speech is expected to degrade for high-frequency regions. The LP residual is known to capture discriminating information for the replay SSD task [10,199–201]. In this context, according to a proposition by Mallat [6], a function $s(t)$ is bounded and k times continuously differentiable with bounded derivatives if

$$\int_{-\infty}^{+\infty} |S(\omega)|(1 + |\omega|^k)d\omega < +\infty, \quad (4.40)$$

where $S(\omega) = \mathcal{F}\{s(t)\} \in L^1(\mathbb{R})$, under the assumption of a Sobolev space. It is known that the decay of spectrum $|S(\omega)|$ of a signal $s(t)$ depends on the worst singular behaviour [6]. For example, in replay speech, the replay noise has sudden discontinuities which are absent in genuine speech. Hence, the spectrum of replay speech is decaying in nature, which is the discriminative acoustic cue For the SSD task by the LFRCC feature set [5]. However, the work reported in [5] proposed LFRCC feature set on the ASVSpooof 2017 v2.0 dataset, whereas this thesis extends it to an *order optimized* LFRCC feature set, and on the simulated replay SSD task on the ASVSpooof 2019 PA dataset.

For the SSD task, the frequency spacing at higher frequencies is sparse (such as in Mel frequency warping). Therefore, to consider the effect of replay mechanism on higher frequency regions, we consider linear frequency scale in this work and exploit linear subband energies. Furthermore, we exploit the recently proposed Linear Frequency Residual Cepstral Coefficients (LFRCC) feature set for the ASV spooof 2019 PA dataset. Unlike [5], we have analyzed the effect of LP order on the residual and hence, proposed the optimized LFRCC feature set.

4.3.1 Linear Prediction (LP)

LP is one of the most powerful methods to analyze speech signals, especially in speech coding for wireless communication services. LP coefficients for speech implicitly represent the time-varying vocal tract area function. It is an iterative method to estimate the current sample of speech $\tilde{s}(n)$, using the past p speech samples because Linear Prediction Coefficients (LPCs) (denoted by $\{\alpha_k\}_{k \in [1,p]}$) capture implicitly the time-varying area function of vocal tract during speech production, where p represents predictor memory [202]. Mathematically, this is rep-

resented as [10]:

$$\tilde{s}(n) = - \sum_{k=1}^p \alpha_k s(n-k). \quad (4.41)$$

With respect to the source-filter model of speech production, the LP method decomposes speech signal into two components: LPCs (representing the vocal tract system using LP filter), and the LP residual (representing the speech excitation source) [203]. By minimizing the squared differences between the actual speech samples and the linear predicted speech samples, a unique set of predictor coefficients can be obtained. The prediction error is called the LP *residual*, as shown in equation ((4.42)). It carries the excitation source component of the speech, and it is given by [204]:

$$r(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^p \alpha_k s(n-k). \quad (4.42)$$

The LP residual is obtained by the all-pole inverse filter $A(z)$, which is mathematically represented in eq. (4.43):

$$A(z) = 1 + \sum_{k=1}^p \alpha_k z^{-k}. \quad (4.43)$$

Furthermore, replayed speech signal ($s_r(n)$) can be expressed as a distributive property of convolution under the assumption of a Linear Time-Invariant (LTI) system, i.e.,

$$s_r(n) = \left[- \sum_{k=1}^p \alpha_k s(n-k) + r(n) \right] * h_r(n), \quad (4.44)$$

where $*$ indicates the convolution operation, and $h_r(n)$ is the impulse response of the playback device used for the replay attack. Notably, the information carried by the LP residual also depends on the LP order, p . A large value of order will lead to good prediction of speech and, hence, lower error (i.e., LP residual).

4.3.2 Proposed Optimized LFRCC

For the SSD task, our aim is not to have a good prediction of speech samples, but rather to exploit the residual at an order optimally suited for the replay SSD task. In particular, for good prediction of speech, the LP residual must not contain any dependencies in the sequence of samples of the LP residual, and thus, the LP residual should be noise-like and hence, its spectrum is expected to be maximally flat. On the other hand, for replay SSD, the LP residual is expected to experi-

ence a significant decay in higher frequencies and thus, it will have different LP order. This is the novel aspect of our work. Figure 4.21 shows waterfall plot of

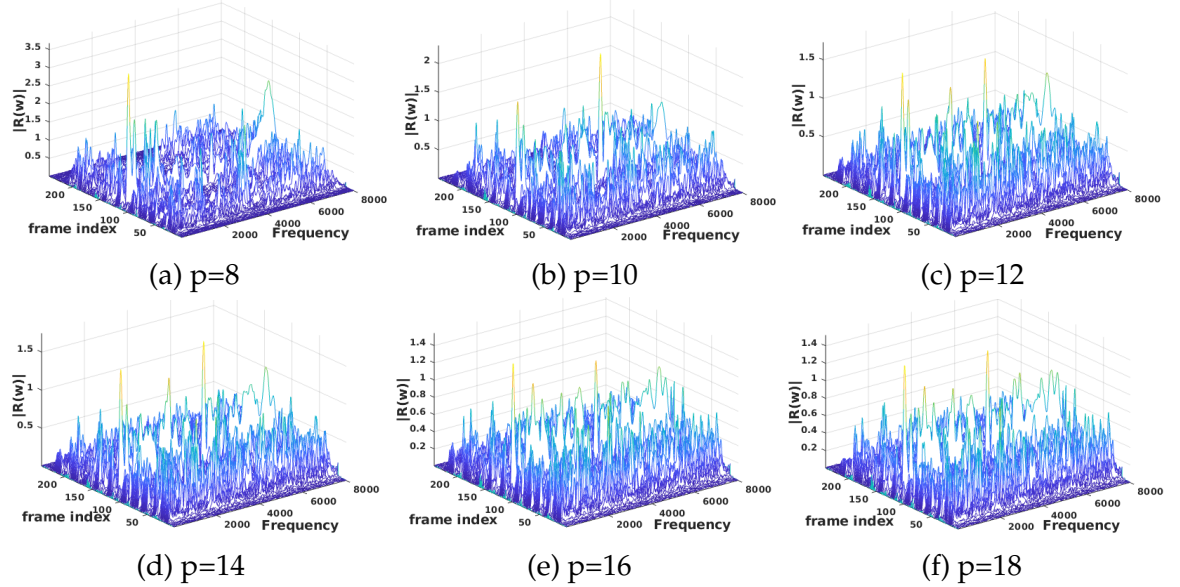


Figure 4.21: Plots of framewise magnitude spectrum $|R(\omega)|$ of the LP residual of genuine speech, for different values of LP order p .

the magnitude spectrum of the residual for varying order, p . It can be observed that the plot has highest $|R(\omega)|$ for $p = 8$, where $R(\omega) = \mathcal{F}\{r(n)\}$, where $\mathcal{F}\{\cdot\}$ represents the Fourier transform. For Figure 4.21, the speech sample taken into consideration had 16 kHz sampling frequency (F_s). This means that the optimum prediction would be achieved at $((F_s/1000) + 2)$, i.e., at order $p = 18$ [196]. However, for exploiting source-based information For the SSD task, the residual should have more information. Hence, $p = 8$ gives the optimal order for the replay SSD task. In addition, the Table 4.16 shows Log Spectral Distance (LSD) between residuals of different LP orders. The LSD is estimated as [205]:

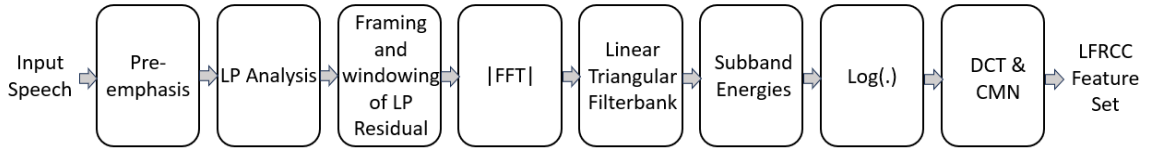


Figure 4.22: Functional Block diagram of LFRCC Feature Extraction. After [5].

$$LSD = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \left[10 \log_{10} \frac{P(\omega)}{\tilde{P}(\omega)} \right]^2 d\omega}, \quad (4.45)$$

where $P(\omega)$ and $\tilde{P}(\omega)$ denote the two power spectra between which the LSD is estimated. The diagonal elements of Table 4.16 are zero because the LSD between

Table 4.16: Log Spectral Distance (LSD) Between the LP Residuals of Speech Signal (with $F_s = 16$ kHz) with Various LP Orders (p).

p	2	4	6	8	10	12	14	16	18	20
2	0	1.35	2.19	2.46	2.64	2.81	2.97	3.06	3.17	3.23
4	1.35	0	1.42	1.74	1.97	2.16	2.34	2.44	2.55	2.61
6	2.19	1.42	0	0.79	1.08	1.31	1.51	1.63	1.75	1.82
8	2.46	1.74	0.79	0	0.63	0.95	1.20	1.33	1.47	1.54
10	2.64	1.97	1.08	0.63	0	0.62	0.94	1.09	1.24	1.32
12	2.81	2.16	1.31	0.95	0.62	0	0.62	0.82	1.00	1.10
14	2.97	2.34	1.51	1.20	0.94	0.62	0	0.47	0.71	0.83
16	3.06	2.44	1.63	1.33	1.09	0.82	0.47	0	0.49	0.65
18	3.17	2.55	1.75	1.47	1.24	1.00	0.71	0.49	0	0.38
20	3.23	2.61	1.82	1.54	1.32	1.10	0.83	0.65	0.38	0

two identical signals is zero. It can be observed that as we move from left to right in the Table 4.16, the LSD keeps on increasing. This also means that the LP order p has a significant effect on the amount of information carried by the LP residual. Furthermore, the experimental results shown in the next Section also confirm our hypothesis that LP order of 8 is *optimal* For the SSD task.

4.3.3 Setup

- **Dataset Used:** The performance of the order optimized LFRCC feature set is evaluated using the ASVSpooof 2019 PA dataset. The details of the dataset are given in subsection 3.3 of Chapter 3.
- **Classifiers Used:**
 - **GMM:** The GMM-based classifier is as explained in subsection 3.4.1 of Chapter 3. In particular, for LFRCC experimentations, the number of mixtures in GMM is taken to be 512.
 - **CNN:** The CNN architecture used in our experiments has 3 convolutional layers (i.e., Conv1, Conv2, and Conv3). After the convolution operation, in order to introduce non-linearity in the neuron output, an activation function is used. In our CNN architecture, we use the Rectified Linear Unit (ReLU) as the activation function. This operation is followed by a pooling layer of kernel size of 3×3 and stride 1 is used. The flattened output is then fed to 2 Fully-Connected (i.e., FC1 and FC2) layers. The output of the final FC2 layer gives us a probabilistic output for classification. The loss function used is binary cross-entropy, and

the optimization algorithm used is gradient descent.

4.3.4 Experimental Results

We present experimental results on ASV spoof 2019 PA dataset using LFRCC feature set For the SSD task. We consider the effect on the EER due to various evalua-

Table 4.17: Effect of LP Order on %EER for LFRCC Features on the ASVSpooof 2019 PA Dataset.

Prediction Order (p)	% EER (GMM)		% EER (CNN)	
	Dev	Eval	Dev	Eval
6	6.84	18.21	6.23	16.17
8	6.77	17.30	6.08	15.21
10	6.89	19.53	5.35	16.88
12	7.02	19.80	6.72	17.20
14	7.19	20.63	6.96	18.37
16	7.54	20.42	7.34	19.28
18	8.38	21.84	7.94	20.36
20	9.43	23.49	8.86	21.11
24	10.97	24.82	8.93	21.89

tion factors, such as LP order, and number of subband filters. Table 4.17 shows the effect of LP order on the EER for LFRCC feature set. It can be observed that the best achieved EER is 15.21% on the Eval set using CNN. Furthermore, on GMM, the best achieved EER is 17.30%. Both of these results are obtained when LP order is kept 8, which is hypothesized as optimal (through an analysis as discussed in subsection 4.3.2). To that effect, the LP order is fixed as 8 for the rest of the experiments in this section.

Additional experimental results to observe the impact of subband filters, and dimension of feature vector as shown in Table 4.18. While keeping the LP order

Table 4.18: Effect of Number of Subband Filters on EER.

No. of Subband Filters	%EER (GMM)		%EER (CNN)	
	Dev	Eval	Dev	Eval
40	6.77	17.30	6.08	15.21
60	6.85	19.01	4.87	17.70
80	7.14	20.47	4.16	18.29
100	7.93	18.28	5.21	17.72
120	8.40	16.32	6.78	15.26
140	9.10	16.79	7.53	14.83

as 8, the number of subband filters in the filterbank are varied from 40 to 140. It can be observed that the best performance on the Eval set using GMM as a classifier is 16.32%. This is obtained when the number of subband filters is 120. Furthermore, when CNN is used as the classifier, the best performance of 14.83% EER is observed, when the number of subband filters is 140. These observations indicate that the optimized LP order for replay spoof detection on the ASVSpooof 2019 PA dataset is 8. However, the performance of the countermeasure system is also improved by increasing the number of subband filters, i.e., by increasing the spectral resolution in the frequency domain.

4.4 U-Vector

Replay SSDs have been known to exploit acoustical features, such as spectral, time-domain, cepstral, and excitation source [206]. The spectral differences between genuine and replayed signal (as shown in [207]), can possibly relate to the “richness” of information of the two signals. This richness of information is captured by a joint representation of signals both in the time and frequency domain simultaneously, obtained using Time-Frequency Distributions (TFD). These representations are limited by the *Heisenberg’s uncertainty principle* [6], in the signal processing framework (detailed proof given in Appendix E). The TFD functions are used to represent energy spectral density of signals jointly in time and frequency-domains [208–210], and the richness of information, is represented by the area of the Heisenberg’s box [211]. This area is characterized mathematically by Time-Bandwidth Product (TBP). A large value of TBP represents more information content of the signal in the duration under consideration. TBP is an indication of characteristics of a stochastic (random) process that has produced the signal, i.e., its sample functions under consideration. The value of TBP can be associated with the number of sample points needed to generate the distribution of the stochastic process [170, 211]. In this section, we propose the *uncertainty vector* (U-Vector), which is based on capturing the richness of information in the signal using Heisenberg’s uncertainty principle. Subsequently, the other two feature vectors, namely, t -vector and ω -vector are also introduced in this work representing the time and frequency-related components of the speech signal, respectively. These feature vectors, unlike the other handcrafted features are easy to reproduce, thereby acting as an effective candidate for practical deployment of SSD for ASV systems.

4.4.1 Time-Bandwidth Product (TBP)

Uncertainty measures the randomness of the process by which the signal is generated. A naturally uttered speech signal has uncertainty because no two similar-sounding speech signals are exactly the same, i.e., there will be *variance*. Hence, there is an intrinsic variance to it created naturally due to the non-linear nature of speech production mechanism. However, in the case of a replayed signal, the variance to the replayed version is due to the added acoustic noise of the playback environment and the playback device, which are considered (or modelled) to be Linear Time-Invariant (LTI) systems. Therefore, there exists difference between a genuine and replayed speech w.r.t. the TBP. Hence, TBP could help by distinguishing signals as genuine or spoof, based on the information content of the signals. The significance of this study is that unlike conventional spoof detection techniques based on cepstral features, TBP-based features gives comparable performance without going into cepstral-domain.

All practical non-stationary signals are finite in duration and have finite bandwidth. Let $x(t)$ be a signal having Fourier transform, $X(\omega) = F\{x(t)\}$. If $X(\omega)$ was to decay quickly in high frequency region, then it means that $x(t)$ must have *regular* time variations. This means that the energy of $x(t)$ has to be spread over a longer range [6]. The time spread can be restricted by doing the following operation given by:

$$x_s(t) = x\left(\frac{t}{s}\right), \quad (4.46)$$

where the scaling factor, $s < 1$. Similarly, its Fourier transform can be given by using time-scaling property [212]:

$$X_s(\omega) = |s|X(s\omega). \quad (4.47)$$

Eq. (4.47) shows that the Fourier transform is dilated by $\frac{1}{s}$. Thus, it shows that being able to gain time localization counter-effects in the frequency domain and vice-versa [212].

The energy spread in the time and the frequency domain is restricted by Heisenberg's uncertainty principle in the signal processing framework [?]. This principle is originally developed in quantum mechanics literature, where it is not possible to find the *position* and *momentum* of a particle simultaneously [213]. Hence, the average location of a particle (signal) $x(t) \in L^2(R)$, (i.e., Hilbert space of finite

energy signals) is given by:

$$\hat{t} = \int_{-\infty}^{\infty} \frac{1}{\|x\|^2} t |x(t)|^2 dt, \quad (4.48)$$

and the average momentum is given by:

$$\hat{\omega} = \int_{-\infty}^{\infty} \frac{1}{2\pi\|x\|^2} \omega |X(\omega)|^2 d\omega, \quad (4.49)$$

where $\|x\|$ represents the L^2 norm of the signal $x(t)$ in Hilbert space. The eq. (4.49) is also called as effective bandwidth as defined by Gabor [214]. The variances around these averages, i.e., σ_t^2 and σ_ω^2 denote the uncertainty in specifying the location and momentum of the particle, respectively. In particular,

$$\sigma_t^2 = \int_{-\infty}^{\infty} \frac{1}{\|x\|^2} (t - \hat{t})^2 |x(t)|^2 dt, \quad (4.50)$$

and the average momentum is given by:

$$\sigma_\omega^2 = \int_{-\infty}^{\infty} \frac{1}{2\pi\|x\|^2} (\omega - \hat{\omega})^2 |X(\omega)|^2 d\omega. \quad (4.51)$$

From eq. (4.46) and eq. (4.47), it can be said that the *expansion* of the signal in one-domain corresponds to *compression* in another-domain. Therefore, the product $\sigma_t^2 \sigma_\omega^2$ is constant and is called a s Time-Bandwidth Product (TBP). Furthermore, TBP also represents the area of Heisenberg's box, which describes the "richness" of information in the given segment of the signal under consideration [144, 170, 215]. This is because the total number of samples required to represent the signal, i.e., the number of degrees of freedom in the signal, is equal to the value of TBP [216]. The value of TBP is found to be a constant for non-stationary signals, such as speech signals. From eq. (4.46) and eq. (4.47), it can be said that the *expansion* of the signal in one-domain corresponds to *compression* in another-domain. Thus, there exists an *inverse* relation between the spread of the signal in either of the domains. Therefore, the product $\sigma_t^2 \sigma_\omega^2$ is constant and is called as the Time-Bandwidth Product (TBP). Furthermore, TBP also represents the area of Heisenberg's box, which describes the "richness" of information in the given segment of signal under consideration [144, 170, 215]. This is because the total number of samples required to represent the signal, i.e., the number of degrees of freedom in the signal, is equal to the value of TBP [216]. The value of TBP is found to be a constant for non-stationary signals, such as speech signals.

From the definitions of variances in time and frequency of a signal in eq. (4.50) and eq. (4.51), the lower bound on TBP is given by the uncertainty principle [6]. In particular, *the time variance* (σ_t^2) *and the frequency variance* (σ_ω^2) *of a signal* $x(t) \in L^2(\mathbb{R})$ *satisfy* (proof is given in Appendix E):

$$\sigma_t^2 \sigma_\omega^2 \geq \frac{1}{4}. \quad (4.52)$$

This means that the smallest possible area covered by Heisenberg's box is 0.25, which is obtained when the signal under consideration is Gaussian in nature [6]. In this work, the values of σ_t^2 , σ_ω^2 , and the product $\sigma_t^2 \sigma_\omega^2$ are used to extract *discriminative* features for the replay SSD task.

4.4.2 U-Vector Feature Extraction

The feature extraction for the replay SSD task is based on the hypothesis that both genuine and spoof utterances possess differences in their spectral energy density and hence, should possess different TBP. This difference in TBP is used as a discriminative acoustic cue for the replay SSD task. Figure 4.23 shows a schematic

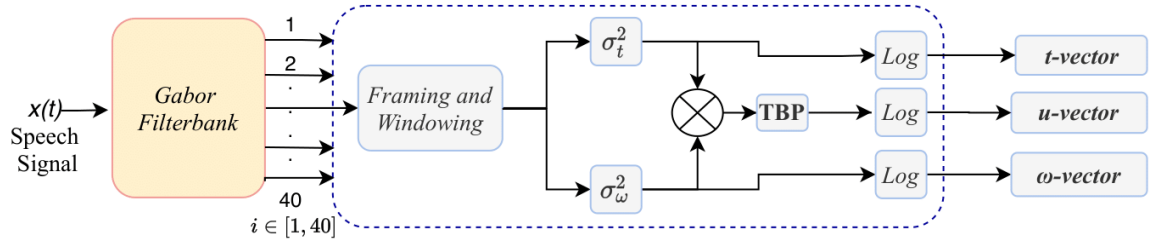


Figure 4.23: Functional block diagram of the proposed t -vector, ω -vector, and u -vector.

representation of the feature extraction procedure of u -vector, t -vector, and ω -vector. Any speech signal $x(t)$ is multi-component in nature and hence, it is first passed through a filterbank. Here, the signal is bandpass filtered using the Gabor filterbank to get several subband signals $x_i(t)$, where $i \in [1, 40]$ [217]. A linearly-spaced Gabor filterbank having 40 subband filters is used because of its optimal time and frequency resolution [6,10]. Each of the subband output signals is frame-blocked with a window size of 70 ms (experimentally chosen), and window shift duration of 10 ms. For each of these frames, both σ_t^2 and σ_ω^2 are computed using eq. (5) and eq. (6) and hence, three different vector representations of the input speech signal are obtained. Logarithm operation is performed on σ_t^2 and σ_ω^2 to give t -vector and ω -vector representations of the speech signal. Similarly, logarithm on the product $\sigma_t^2 \sigma_\omega^2$ results in u -vector, where u represents the uncertainty

of the input speech signal. Here, Discrete Cosine Transform (DCT) was found to degrade the performance of the proposed feature sets. Hence, all the subsequent operations, namely, Cepstral Mean and Variance Normalization (CMVN), velocity, and acceleration coefficients, are not considered here for analysis. We have

$$\log(\sigma_t^2 \sigma_\omega^2) = \log(\sigma_t^2) + \log(\sigma_\omega^2), \quad (4.53)$$

$$u - \text{vector} = t - \text{vector} + \omega - \text{vector}. \quad (4.54)$$

Equation (4.54) describes the relation between u -vector, t -vector, and ω -vector.

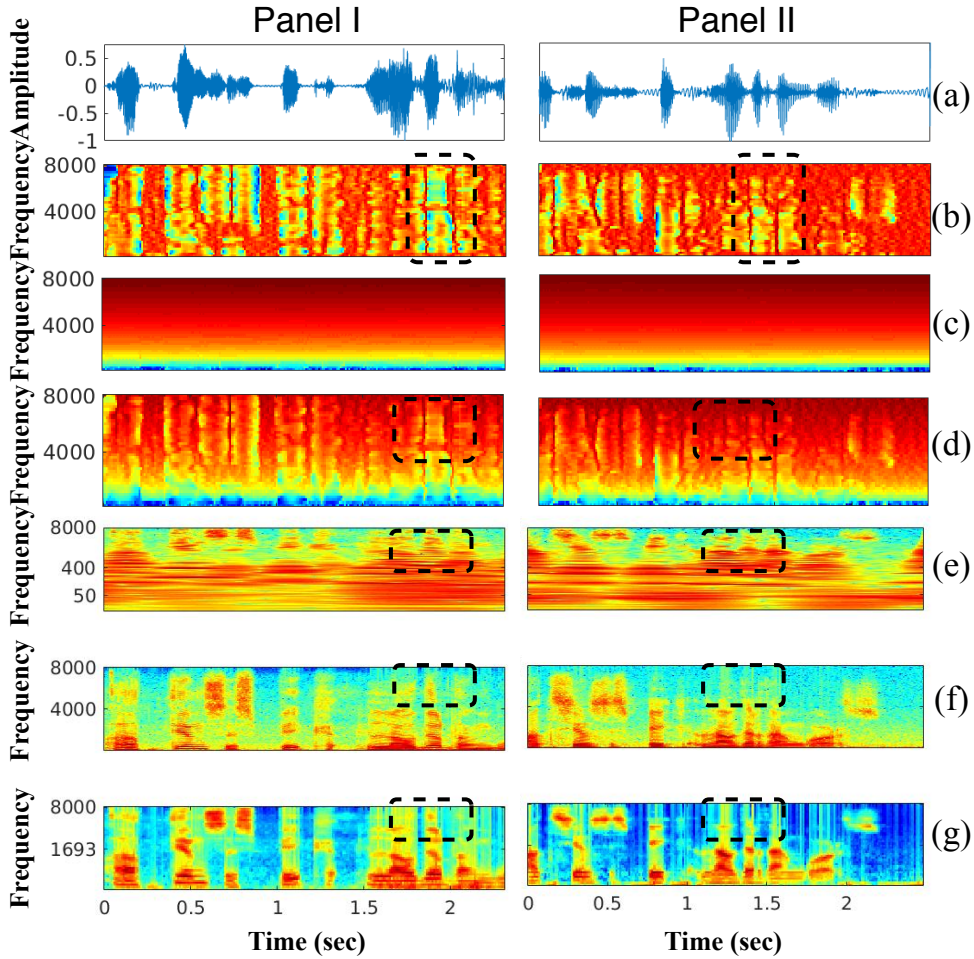


Figure 4.24: Representation of a genuine speech signal (Panel-I) *vs.* representation of a spoof speech signal (Panel-II): (a) speech signal, (b) t -gram, (c) ω -gram, (d) u -gram, (e) CQT-gram, (f) spectrogram, and (g) Mel-spectrogram. Ordinate values of the Panel-II subplots are similar to that of Panel-I. Abscissa values of 4.24 (a)-(f) are similar to that of Figure 4.24 (g).

The time-frequency representation of u -vector, t -vector, and ω -vector for an utterance is denoted by u -gram, t -gram, and ω -gram, respectively. Figure 4.24 shows the effect of the replay channel on the genuine speech signal. Panel-I and

Panel-II show the analysis on genuine and spoof speech utterance, respectively. It can be observed that, amongst t -gram and ω -gram, t -gram shows more *discriminating* ability for genuine *vs.* replay spoof speech signal. The t -vector corresponds to the variance of the given frame-blocked signal for a particular subband filter output. The lesser the signal spread, the lesser would be its variance and hence, the more localized the signal is. The replay spoof signal consists of additional components coming from the replay channel and the reverberation coming from the acoustic environment. Because of these components and distortions, the replay signal appears to be more spread in the time domain as compared to its genuine counterpart. This leads to an increase in the time variance (σ_t^2). This fact is clearly visible in Figure 4.24(b), as the t -gram for genuine speech signal has more black-coloured regions representing lesser σ_t^2 than that for the replay speech signal. This shows that t -vector is effective in distinguishing between sharp and gradual changes along the speech segment. The value of the σ_ω^2 , on the other hand, increases with an increase in the center frequencies of the subband filters. ω -gram for genuine and spoof speech utterances appears to be the same. It could also be observed that this might be the reason that ω -vector could not perform well For the SSD task. It is also observed that the value in ω -vector is dominant in the high-frequency regions and much lower in the low-frequency regions (as represented by the black-colored regions). In addition, speaker-specific cues are known to exist more in the high-frequency regions as compared to the low-frequency regions [218–221]. This inherent attribute of ω -vector eliminates the need for highpass filtering at the pre-processing step for effective SSD task. Thus, higher values of σ_ω^2 and its ability to emphasize high-frequency regions effectively compensate the lower values of σ_t^2 and its ability to easily identify sharp regions make their summation, i.e., the u -vector, superior to σ_t^2 and σ_ω^2 . However, in ω -gram the low-frequency regions are quite suppressed, and therefore, this can be the possible reason as to why considerably large window size 70 ms gives better %EER, as shown in the results presented in subsection 4.4.4. Hence, large window size is a direct implication of the constraints imposed by Heisenberg’s uncertainty principle on the low-frequency components of the speech signal. We also included Constant Q Transform (CQT)-gram, spectrogram, and Mel-spectrogram for comparison with the proposed representations. The highlighted region in Figure 4.24(b), 4.24(d), 4.24(e), 4.24(f), and 4.24(g) shows differences between the genuine, and spoof parts of the same region of an utterance. It can be observed from Figure 4.24(d), that u -vector shows more discriminative ability in the high-frequency region. Moreover, u -gram also shows comparatively more

spectral fading in high-frequency region, which is a peculiar feature of a replayed speech utterance [222].

4.4.3 Setup

- **Dataset Used:** For evaluating the performance of u -vector, t -vector, and ω -vector, the ASVSpooF 2017 v2.0 dataset is used whose details are given in subsection 3.3 of Chapter 3.
- **Classifiers Used:** The classifiers used were GMM (with 512 mixtures), CNN, and LCNN.

4.4.4 Experimental Results

Figure 4.25 shows the variation of % EER obtained for various window sizes (in ms) for the u -vector performed on Dev set of the ASVSpooF 2017 V2.0 dataset. Amongst these, the window size that gave the best % EER is chosen (70 ms) and is used for finding evaluation results. Table 4.19 shows the performance of the

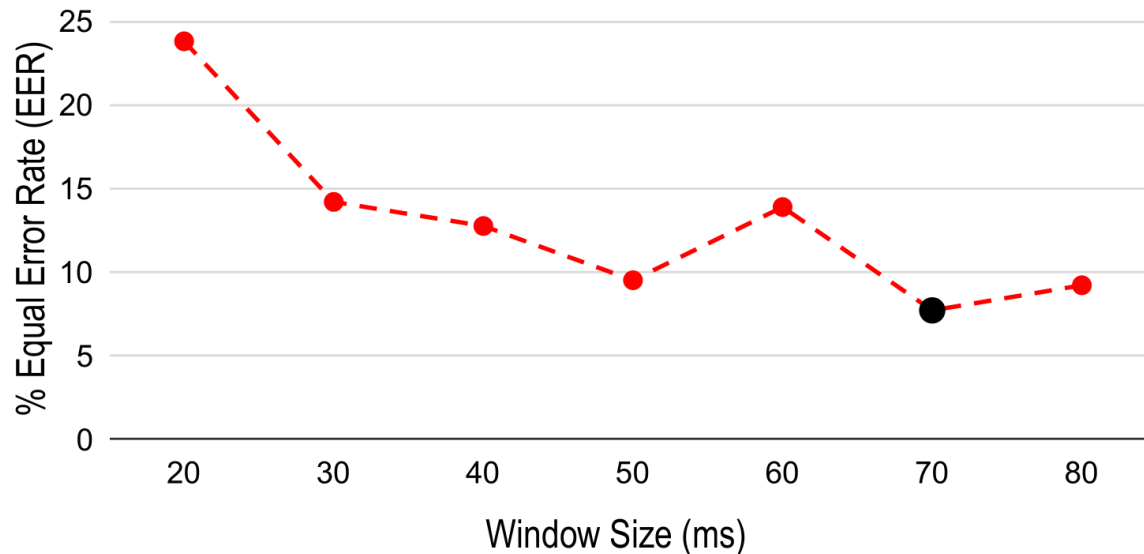


Figure 4.25: Variation of % EER *vs.* window size (ms) for u -vector.

feature sets. The symbol \oplus denotes the score-level fusion. The results of the proposed feature set are compared against the baseline CQCC that gives 12.27% and 18.81% EER on Dev and Eval sets, respectively. Similarly, the performance of the MFCC and LFCC feature sets is also shown in Table 4.19. GMM classifier is used unless stated otherwise. From Table 4.19, it can be observed that the performance of u -vector and t -vector is superior to the baseline CQCC ($S1$), MFCC ($S2$),

Table 4.19: Results on Dev and Eval sets for systems trained on GMM, CNN, and LCNN.

Feature Set	% EER	
	Dev	Eval
CQCC (S1)	12.27	18.81
MFCC (S2)	22.39	25.34
LFCC (S3)	17.30	17.00
t -vector	7.63	15.58
ω -vector	26.37	33.62
u -vector (S4)	7.72	13.53
u -vector-CNN	14.73	21.10
u -vector-LCNN	24.21	26.88
u -vector \oplus CQCC	7.47	12.48
u -vector \oplus MFCC	6.82	12.11
u -vector \oplus LFCC	7.48	12.02
t -vector \oplus ω -vector	6.15	14.94
S1 \oplus S2 \oplus S3	9.95	14.30
S1 \oplus S2 \oplus S3 \oplus S4	6.66	11.99

and LFCC (S3) feature sets. u -vector (40 linearly-spaced Gabor subband filters) gives an absolute reduction of 4.55% and 5.28% in % EER on Dev and Eval sets, respectively, when compared to the baseline system. Furthermore, experiments performed for Mel-spaced Gabor subband gave results as 26.03% and 26.9% on Dev and Eval sets, respectively. Since u -vector emphasizes more in the higher frequency regions, placing fewer subband filters in high frequency regions as in case of Mel-spaced subband filters, is the possible reason for these results. Hence, the Gabor filters are linearly-spaced in this work. In addition, t -vector gave an absolute reduction of 4.64% and 3.23% in EER on Dev and Eval sets, respectively, when compared to the baseline system. Table 4.19 also shows that the GMM classifier gives superior results when compared with CNN and LCNN classifiers, as these deep learning classifiers require a large amount of training data. Score-level fusion of u -vector with LFCC gave the best EER of 8.77% and 12.02% on Dev and Eval sets, respectively, suggesting that both the feature sets encapsulate complementary information. The result is further improved on fusion of systems S1, S2, S3, and S4 over the fusion of systems S1, S2, and S3.

Figure 4.26 shows the DET plots for the CQCC, MFCC, LFCC, u -vector, and score-level fusion of all the feature sets. u -vector-based SSD system shows the best performance amongst all individual systems. The performance of the score-level fusion of all the system is shown by the black dotted line, which gives EER of 11.99%.

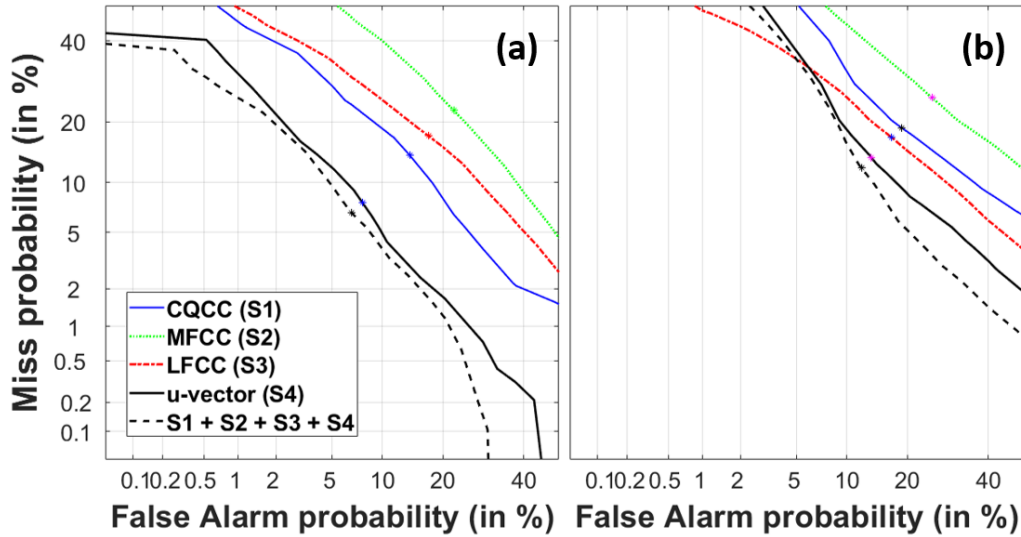


Figure 4.26: DET curves for replay SSD system. The individual DET curves for u -vector (proposed), t -vector, ω -vector, CQCC, MFCC, LFCC on (a) Dev set, and (b) Eval set.

4.5 Chapter Summary

This chapter proposed three feature sets for the replay SSD task, namely, CFCCIF-QESA, optimized LFRCC, and u -vector. The CFCCIF-QESA feature set incorporates the quadrature phase along the in-phase component of the speech signal, which helps to capture additional information of the signal. The significance of incorporating quadrature phase is shown through mutual information (MI)-based analysis. Furthermore, the performance of the CFCCIF-QESA feature set is validated on various datasets, such as ASVSpooof 2015, ASVSpooof 2017 v2.0, ASVSpooof 2019, VSDC, ReMASC, and BTAS 2016, and it is found that the addition of quadrature phase enhances the performance of the SSD system, as compared to the existing CFCCIF-ESA feature set. Additionally, the optimized LFRCC feature set is evaluated for replay SSD on the ASVSpooof 2019 PA dataset. It is found that the optimal order for linear prediction is not the same as that needed for the replay SSD task. Furthermore, Heisenberg’s uncertainty principle in the signal processing framework is exploited to design the u -vector feature set. The u -vector further results from two feature sets, namely, t -vector and ω -vector. The performance of each of these feature sets is evaluated on the ASVSpooof 2017 v2.0 dataset.

When considering the real-world scenario where an attacker is an external entity who is free to select any strategy of generating the spoofed signal, the reliability of the present SSD systems on a specific attack type prevents them from being built as a generalized SSD system. This is mostly because the current SSD systems

rely on spoofed signal characteristics to determine whether a speech utterance is spoofed. Therefore, by utilizing the characteristics of genuine (live) speech rather than spoofed speech, VLD systems represent a promising step towards addressing this issue. In this context, the next chapter is aimed to design efficient VLD system.

CHAPTER 5

Features for Voice Liveness Detection (VLD)

5.1 Introduction

In the last chapter, we saw the implications of real wavelet transform (i.e., auditory transform) to develop the CFCCIF-QESA feature set for the replay SSD task. In this chapter,¹ we present the application of analytic wavelet transform (using various analytic wavelets) for the VLD task. In particular, this chapter is based on the VLD task of detecting live speech using *pop noise* as the discriminating acoustic cue, where microphones have the ability to capture the effect of the breath in the form of pop noise generated from live speech [27, 85, 86]. Pop noise is a common distortion in live speech, occurring due to the proximity of the live speaker's mouth with the microphone [27]. During natural (live) speech production, the

¹This Chapter is based on the following publications:

- **Priyanka Gupta**, and Hemant A. Patil, "Voice Liveness Detection Using Morse Wavelet Transform", submitted in Computer, Speech & Language, Elsevier, 2023, 31 pages.
- **Priyanka Gupta**, and Hemant A. Patil, "Significance of Distance on Pop Noise for Voice Liveness Detection," in International Conference on Speech and Computer (SPECOM), S. R. Mahadeva Prasanna, Alexey Karpov, K. Samudravijaya, and Shyam S. Agrawal (Eds.), Lecture Notes in Computer Science (LNCS), vol 13721, pp. 226-237, 2022, Springer.
- **Priyanka Gupta**, Siddhant Gupta and Hemant A. Patil, "Voice Liveness Detection using Bump Wavelet with CNN," in International Conference on Pattern Recognition and Machine Intelligence (PReMI), Lecture Notes in Computer Science (LNCS), 2021, Springer.
- **Priyanka Gupta**, and Hemant A. Patil, "Effect of Speaker-Microphone Proximity on Pop Noise: Continuous Wavelet Transform-Based Approach" in the 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, pp. 110-114, Dec. 11-14, 2022.
- **Priyanka Gupta**, Piyushkumar K. Chodingala and Hemant A. Patil "Morlet Wavelet-Based Voice Liveness Detection using Convolutional Neural Network", in 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, pp. 100-104, 29 Aug. -02 Sept., 2022.
- **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil, "Morse Wavelet Features for Pop Noise Detection," in 2022 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, July 11-15, 2022, pp. 1-5.

airflow from the lungs causes the vocal folds to vibrate (during voiced sounds), which excites the vocal tract system (that is modelled as a cascade of 2nd order resonators). The airflow, after passing through the cascade of 2nd order resonators, reaches the mouth, where the *place* and *manner* of articulation depend on the type of phoneme being uttered, causing a burst of air from the lips. If the sound is captured by the microphone at a small distance from the speaker, the microphone captures the speech along with energy released due to the friction between the lips as bursts of airflow, which is termed as *pop noise*. The energy of pop noise is expected to decrease with distance (also observed experimentally in this thesis work). This *inverse* relationship between distance and the energy of pop noise can be used to detect replay attacks (assuming the attacker records live speech from a considerably large distance, discreetly from the speaker). Therefore, the inability of the attacker to place the recording device near the speaker leads to the recording of a diminished (or no) pop noise. Furthermore, it is known that on playing the speech with the help of a playback device (or loudspeaker), the loudspeaker fails to reproduce the pop noise [89, 90]. Therefore, irrespective of the algorithm used for generating the spoofed signal, during the mounting of the spoofing attack, pop noise is weakly reproduced by the loudspeaker. This makes pop noise an important acoustic cue to distinguish live speech from spoofed speech played using loudspeaker devices. This has led to research on VLD focused on pop noise [27, 87, 88, 101]. So far, the research in this direction is only on the rise, and the majority of the existing approaches use features with linear resolution in the frequency-domain, such as STFT.

5.2 CWT-Based Approach

Given that STFT has a *fixed* resolution in time and frequency-domains [6], we make use of the Continuous Wavelet Transform (CWT), which has improved frequency resolution at lower frequencies, as shown in Figure 5.1. The analytic CWT of a real-valued signal $x(t)$ is denoted by $W_a x(a, b)$, and is given by [6]

$$\begin{aligned} W_a x(a, b) &= \langle x(t), \psi_{a,b}(t) \rangle, \\ &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt, \end{aligned} \quad (5.1)$$

where $\langle \cdot, \cdot \rangle$ indicates inner product operation to compute wavelet coefficients, and $*$ denotes complex conjugate. The dilation (scaling) coefficient is denoted by a , and translational (positional) coefficients are denoted by b . It should be noted

that the key difference between eq. (4.31) and eq. (5.1) lies in the use of real and analytic wavelets, respectively.

Given that the pop noise exists in low-frequency regions, CWT-based proposed approach effectively captures distinguishing acoustic cues between live and non-live speech. The CWT-based time-frequency representations are known as *scalograms*, which are computed as $|W_a x(a, b)|^2$ from equation (5.1) [6]. Fur-

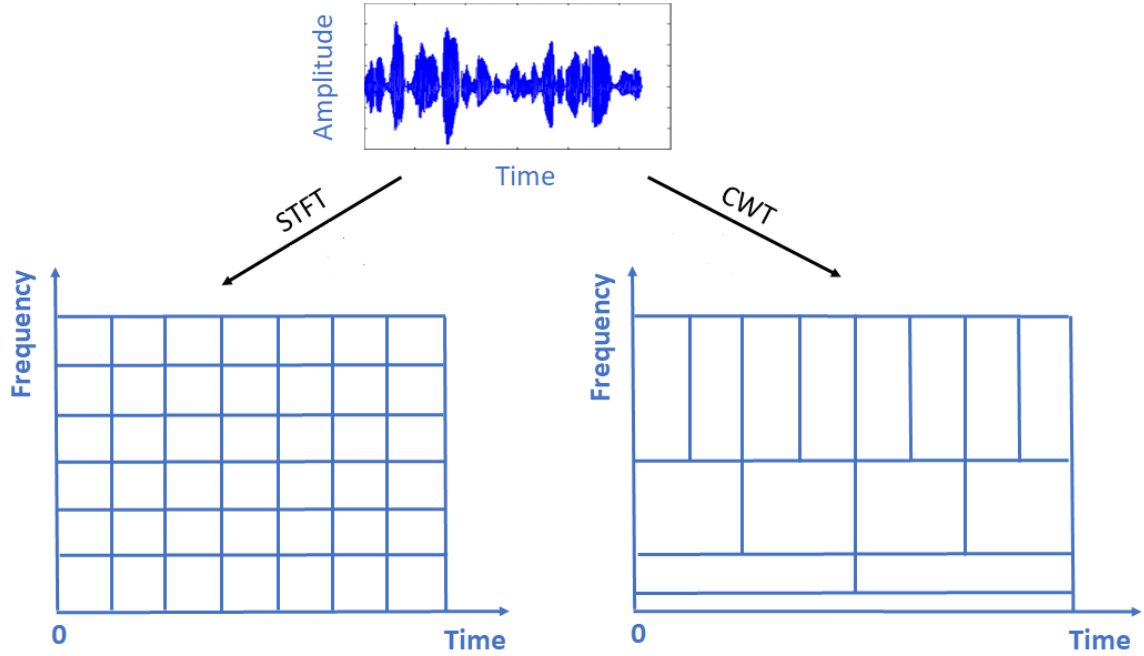


Figure 5.1: Tiling of the Time-frequency plane for STFT *vs.* CWT. After [6].

thermore, CWT can be seen as a filtering approach with lowpass and highpass filtering at various scales, which span the entire time-frequency plane. However, the area of Heisenberg’s box (details are given in Appendix E) remains constant, and its value depends on the type of mother wavelet function chosen for estimating CWT. Furthermore, similar to the STFT, energy conservation is preserved in analytic WT as well, which is expressed mathematically as [6]:

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |W_a x(a, b)|^2 db \frac{da}{a^2}. \quad (5.2)$$

In this work, we include discussions pertaining to analytic wavelets for the VLD task. The motivation to choose analytic wavelets in this work comes from the fact that the signal form used to analyze analytic wavelet properties, comes from the model used to synthesize speech as in [223]. Furthermore, analytic wavelets are known to be used for the analysis of oscillatory signals [214, 224]. These oscillations are encoded in the signal in the form of zero-crossings in a speech signal

and hence, also in its IF [225]. To that effect, IF is detected using analytic function generation. In particular, speech can be modelled as an AM-FM signal, which is an oscillatory signal, whose IF estimation is done using analytic wavelets (e.g., methods such as the classic Delprat’s algorithm for IF estimation [226]). Apart from this, the analytic CWT is robust to noise and stable to perturbations and higher-order modulations, which makes it suitable for feature extraction.

In this chapter, we propose to exploit analytic wavelet-based approaches for pop noise detection for the VLD task. To that effect, three feature sets, namely, Bump wavelet-based, Morlet wavelet-based, and Generalized Morse Wavelet (GMW)-based features are proposed. The organization of the chapter is as shown in Figure 5.2. We begin by discussing analytic Bump wavelet-based features and

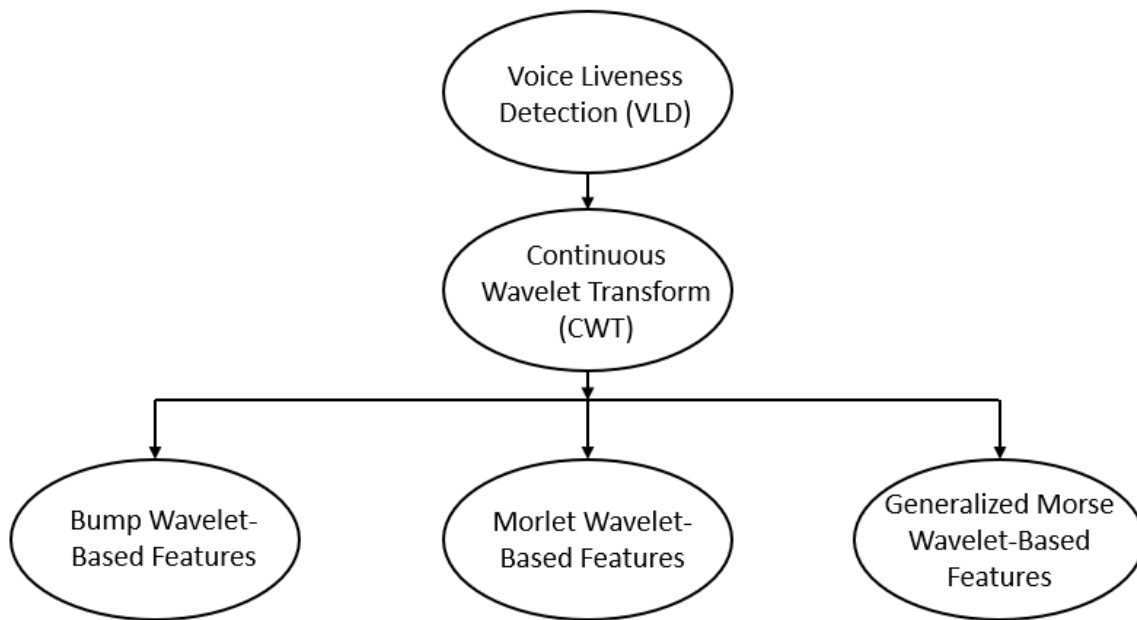


Figure 5.2: Flowchart of the contents of this Chapter w.r.t. the Proposed CWT-Based Features for VLD.

the corresponding analysis of experimental results. This is followed by Morlet wavelet-based features, due to the fact that Morlet wavelet is closely related to human perception (for both hearing and vision) [66]. However, the issue of selecting an appropriate wavelet remains. To alleviate this issue, we further propose GMW-based features, where GMWs are considered as the superfamily of analytic wavelets. Notably, to the best of the author’s knowledge and belief, this is the first work of its kind to report the use of GMWs in speech technologies, more so for liveness detection of speech. To that effect, this chapter presents detailed experimental results on GMWs w.r.t. various evaluation factors, such as varying feature parameters, frequency range, the effect of classifier structure, the effect of

speaker-attacker distance, and the effect of speaker-microphone distance.

5.3 Bump Wavelet-Based Features

5.3.1 Proposed Approach

Live speech contains pop noise, which is caused by sudden burst of human breath on the microphone [26]. Time-frequency representations, such as spectrograms, have been used in the past to locate the pop noise event in speech signals [27, 87]. However, to get better detection of pop noise, we propose CWT-based features using Bump wavelet in this work. The mother Bump wavelet is defined in the frequency-domain and is given by [227]:

$$\Psi(s\omega) = e^{\left(1 - \frac{1}{1 - (s\omega - \sigma)^2 / \sigma^2}\right)} \mathbf{1}_{[(\mu - \sigma)/s, (\mu + \sigma)/s]}, \quad (5.3)$$

where $\mathbf{1}_{[(\mu - \sigma)/s, (\mu + \sigma)/s]}$ is the indicator function over the interval $[(\mu - \sigma)/s, (\mu + \sigma)/s]$. In eq. (5.3), the value of μ lies in the interval $[3, 6]$, whereas the value of σ lies in the interval $[0.1, 1.2]$. For smaller values of σ , we get a wavelet with superior frequency resolution as compared to the time resolution. On the other hand, for larger values of σ , we get a wavelet with superior time resolution as compared to the frequency resolution. For our experiments in this study, we have taken $\mu = 5$ and $\sigma = 0.6$. These values enable us to get optimum resolution in both the time and frequency-domains. The proposed algorithm (as shown in Algorithm 4) uses CWT coefficients corresponding to lower frequency regions, so that the pop noise event is detected efficiently. Figure 5.3 shows the bump wavelet-based scalograms of the word 'thong'. There is a distinct signature of the pop noise in a live speech signal, as shown in Panel I. On the other hand, the pop noise signature is absent for the case of non-live speech as shown in Panel II of Figure 5.3.

In our experiments, the lowest frequency bin is found empirically at 6.4564 Hz, and two consecutive bins are separated by a factor of 1.0718. Therefore, the index k of the bin corresponding to 40 Hz is calculated as:

$$40 = (1.0718)^k * 6.4564. \quad (5.4)$$

Hence, to estimate frequency region below 40 Hz, we get the nearest integer $k = 27$ for the frequency bin corresponding to 41.9537 Hz. This is the region where the pop noise is expected to be located. To that effect, scalogram images are extracted

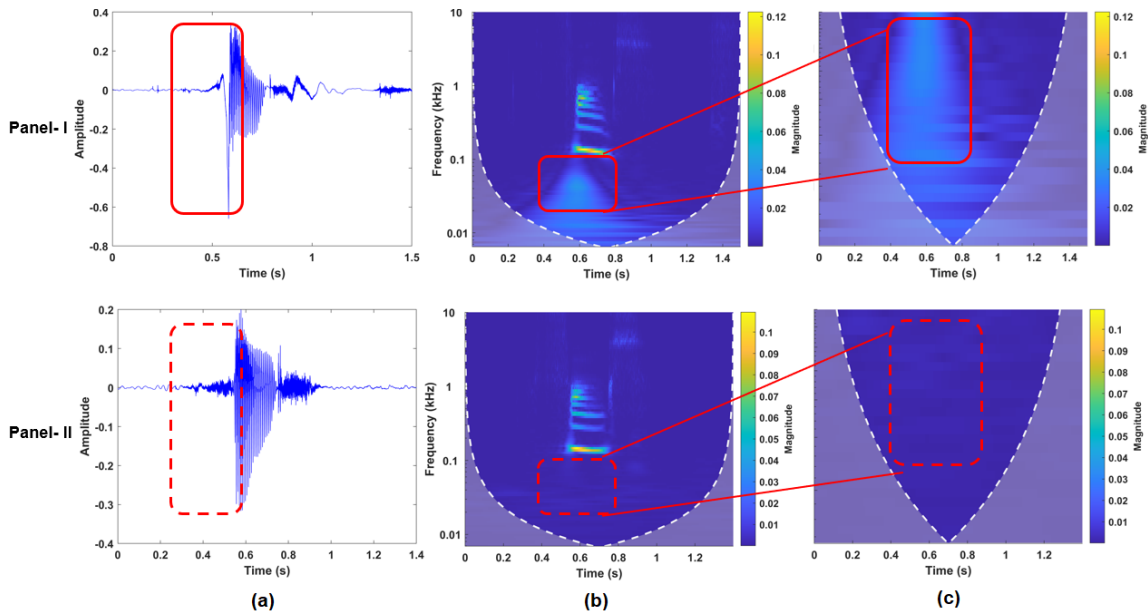


Figure 5.3: Panel I represents the case of presence of pop noise (genuine or live speech). Panel II represents suppressed pop noise (spoofed speech) due to the use of pop filter. (a) Time-domain signal for the word ‘*thong*’, (b) corresponding scalogram, and (c) selected region of scalogram in (b) corresponding low-frequency (0 – 40 Hz). Solid boxes in Panel I indicates the presence of pop noise, while corresponding dotted boxes in Panel II indicates that the pop noise has been suppressed by the use of pop filter.

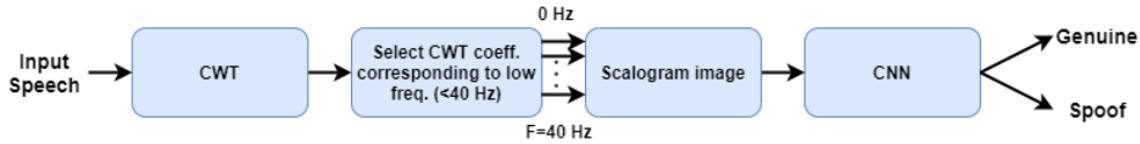


Figure 5.4: Proposed Approach for the VLD task.

corresponding to only those 27 wavelet coefficients, as shown in Algorithm 4. Each scalogram image is of the size $3 \times 512 \times 512$.

5.3.2 Setup

- **Dataset:** The experiments are performed on the POCO dataset. In particular, genuine (live) utterances are taken from the RC-A subset, and the spoofed utterances are taken from the RP-A subset. For the experiments, the utterances are divided into training and Eval sets with details as shown in Table 5.1.

For distance-wise analysis, the RC-B subset of the POCO dataset is used since it consists of a microphone array of 15 microphones. The arrangement of the speaker in front of the microphone array is described in Figure 3.1 in

Algorithm 4 Bump Wavelet-based Feature Extraction for the VLD task.

```
1: procedure FEAT( $x$ ) ▷  $x$  is the speech signal
2:    $w\_name = 'bump'$  ▷ Taking Bump wavelet
3:    $[cwt\_coeffs, F] \leftarrow cwt(x, w\_name)$ 
4:    $Low\_bins \leftarrow find(bins \geq 27)$  ▷ 27 bins correspond to 0 to 40 Hz
5:    $Low\_coeffs \leftarrow cwt\_coeffs(Low\_bins)$ 
6:    $Pop\_energy = [abs(Low\_coeffs)]^2$ 
7:    $rescaled\_energy = rescale(Pop\_energy)$ 
8:    $ind\_image = im2uint8(rescaled\_energy)$  ▷ Convert to 8-bits unsigned integers
9:    $Scalogram = ind2rgb(ind\_image)$  ▷ Convert to RGB image
10: end procedure
```

Table 5.1: Statistics of the POCO Dataset for the Experiments.

Partition	# Utterances	# Speakers	
		Male	Female
Training	13552	26	27
Evaluation	3432	6	7

Chapter 3. Furthermore, due to the arrangement of the microphones, each of the microphones is at a particular distance from the speaker. The details w.r.t. this arrangement along with the distance calculations are shown in Table 3.10 in Chapter 3.

- **Classifier:** For classification, CNN is used as the classifier having the architecture as shown in Figure 5.5, wherein we use three convolution layers and three fully-connected layers. Each convolution layer consists of a 2-D convolution operation with a kernel size of 3×3 . Batch normalization is done for the output of the convolution operation to remove irregularities. Batch normalization normalizes the intermediate outputs of each layer within a batch during training, making the optimization process more stable and faster. By reducing *internal covariate shift*, batch normalization allows for higher learning rates, accelerates convergence, and improves generalization performance, leading to better and more efficient neural network training. An internal covariate shift occurs when there is a change in the input distribution to our network. When the input distribution changes, hidden layers try to learn to adapt to the new distribution. This slows down the training process. If a process slows down, it takes a long time to converge to a global minimum. Further, a max-pooling operation is then performed with a kernel size of 3×3 . Each of the three convolution layers follows the same

structure. The output of the final layer is then *flattened* (i.e., converted to a 1-D representation), and fed to a cascade of three fully-connected layers. The output of the final fully-connected layer is a single numerical value. The final output is then activated using the Sigmoid activation function to convert the value into a probability. All the hidden layers in the network apply ReLU activation function to introduce non-linearities into the output. Since the task is binary classification, binary cross-entropy is used as the loss function. Optimization is done using stochastic gradient descent with a learning rate of 0.0001. The network was trained with a batch size of 32 for a total of 500 epochs.

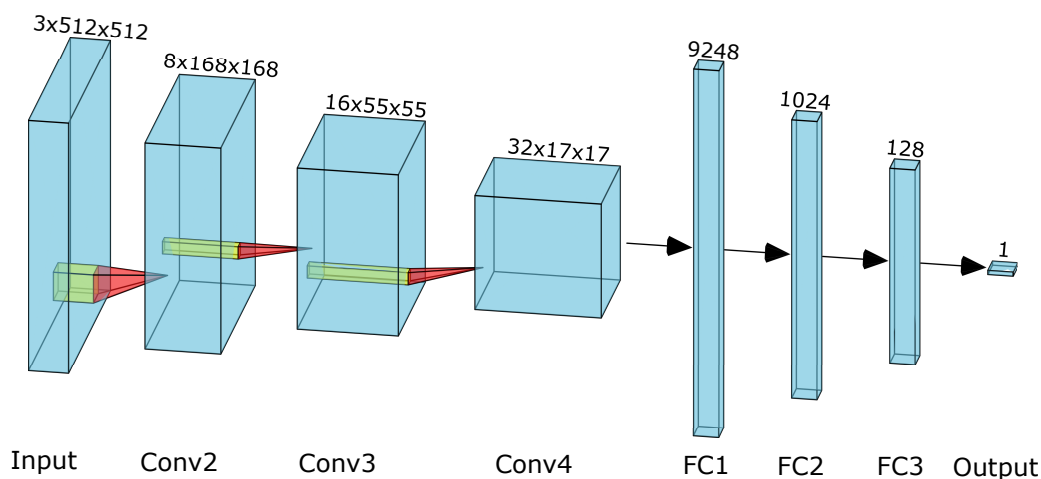


Figure 5.5: The CNN architecture used for classification of the proposed bump wavelet-based scalogram features. After [7].

5.3.3 Speaker-Microphone Distance-Based Analysis

In this subsection, we illustrate the effect of distance between the speaker and the microphone on the energy of pop noise. To that effect, Figure 5.6 depicts 3 cases, where the distance between the microphone and speaker's mouth is taken to be as 5 cm, 5.39 cm, and 6.42 cm, as Panel-I, II, and III, respectively. The utterance spoken is 'dad' taken from the POCO dataset. The time-domain representation of the signal in Figure 5.6 Panel-I, shows that the pop noise is dominantly present as indicated by the red rectangular box. This is the case, where the speaker's mouth is the closest to the microphone (i.e., at 5 cm). Similar observation in Panel-I can be made from the corresponding CWT-based scalogram representations over the entire frequency range till $f_s/2$ (i.e., full frequency) as well as low frequency scalogram representations, where the presence of pop noise energy is highlighted via

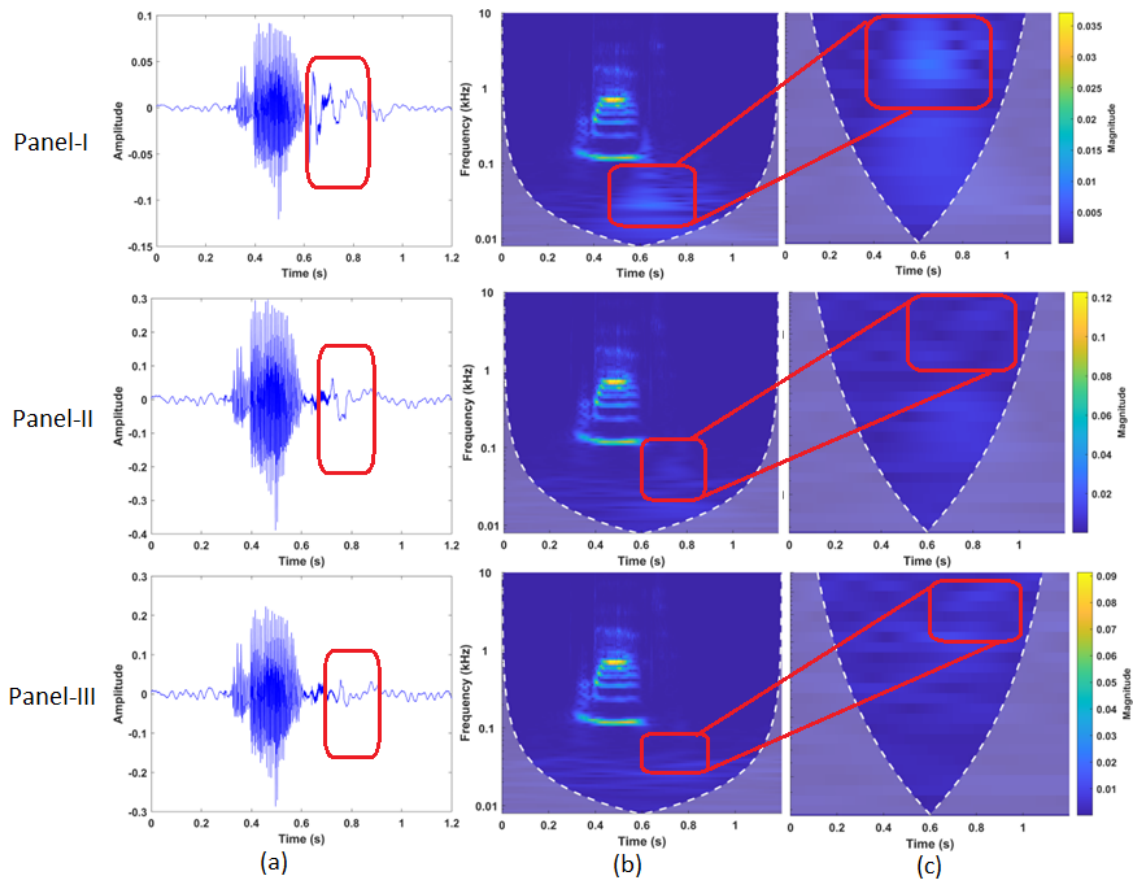


Figure 5.6: Panel I, Panel II, and Panel-III represent the varying distance of the speaker from the microphone, i.e., 5 cm, 5.39 cm, and 6.42 cm, respectively, for (a) time-domain signal for the word 'dad', (b) corresponding scalogram, and (c) selected region of scalogram in (b) corresponding to low-frequency (0 – 40 Hz). Solid boxes in red indicate the presence of pop noise.

red boxes. Following this, the distance between the speaker's mouth and the microphone is increased to 5.39 cm in Panel-II. It can be noted that even though the pop noise is not visible dominantly in the time-domain representation, it is captured by the scalogram-based time-frequency representation. However, the strength of pop noise energy is degraded as compared to Panel-I. Lastly, Panel-III shows the case, when the speaker's mouth is at the farthest distance from the microphone, i.e., 6.42 cm. In this case, the pop noise energy has the lowest strength, for time-domain representation, as well as the scalogram-based representations. The word chosen for this analysis is 'dad', which predominantly contains plosives.

Apart from the distance, the strength of pop noise captured also depends on the type of phonemes present in the word uttered in front of the microphone. Phonemes are produced from a combination of vocal fold and vocal tract articulatory features, where the articulatory features correspond to the vocal fold

state (open or closed), and the tongue position (front, central, and back), i.e., *place* and *manner* of articulation. In this work, we discuss the analysis w.r.t. six types of phonemes, namely, plosives, fricatives, whisper, affricates, nasal, and liquids, where the pop noise strength is estimated using Algorithm 5.

Algorithm 5 Proposed Algorithm for Pop Noise Energy Estimation Using Bump Wavelet for VLD.

```

1: procedure ENERGY_POP( $x$ )                                ▷  $x$  is the speech signal
2:    $w\_name = 'bump'$                                        ▷ Taking Bump wavelet
3:    $[cwt\_coeffs, F] \leftarrow cwt(x, w\_name)$ 
4:    $Low\_bins \leftarrow find(bins \geq 27)$ 
5:    $Low\_coeffs \leftarrow cwt\_coeffs(Low\_bins)$ 
6:    $Pop\_energy = [abs(Low\_coeffs)]^2$ 
7:    $[r, c] = size(Pop\_energy)$ 
8:   for  $i \leftarrow 0 : r$  do
9:      $E\_LF(i) = sum(Pop\_energy(i, :))$ 
10:  end for                                               ▷ Each row of E_LF has energy for 1 frequency bin
11:   $Emean = mean(E\_LF)$ 
12: end procedure

```

To that effect, Figure 5.7 shows the various cases of phoneme categories and the effect of distance on the strength of the pop noise. The trend of pop noise energy is shown with the help of the dotted lines, also referred to as ‘trendlines’. The trendline equations are estimated using the in-built exponential trendline in Microsoft Excel. In particular, for the case of plosives (as shown in Figure 5.7 (a)), the Bump wavelet-based method shows trendline with the equation $y = 0.44e^{-0.233x}$, and the baseline STFT-based method shows trendline with the equation $y = 0.25e^{-0.204x}$. Therefore, we can say that the exponential trendline for the Bump wavelet-based method shows relatively *more rapid* decay as compared to the trendline of STFT. An efficient VLD system should be able to capture more pop noise at smaller distances; and it should also fail to capture pop noise at sufficiently larger distances, thereby having a more rapidly decaying trendline. Therefore, Figure 5.7 (a) shows that the proposed Bump wavelet-based method of pop noise detection is more suited for the VLD task, as compared to the traditional STFT-based method. Similar behaviour of trendlines can be observed in the cases of fricatives, whisper, and affricates, as shown in Figure 5.7 (b), (c), and (d). Hence, the suitability of wavelet-based approach is further enforced in these phonemes, as compared to STFT-based method. However, for nasal and liquids, we see an almost constant-like trendline, which shows that the distance does not affect the pop noise energies in nasals and liquids. Given that liquids are semi-vowels [10], they have very less or no pop noise, because the nasal cavity is large

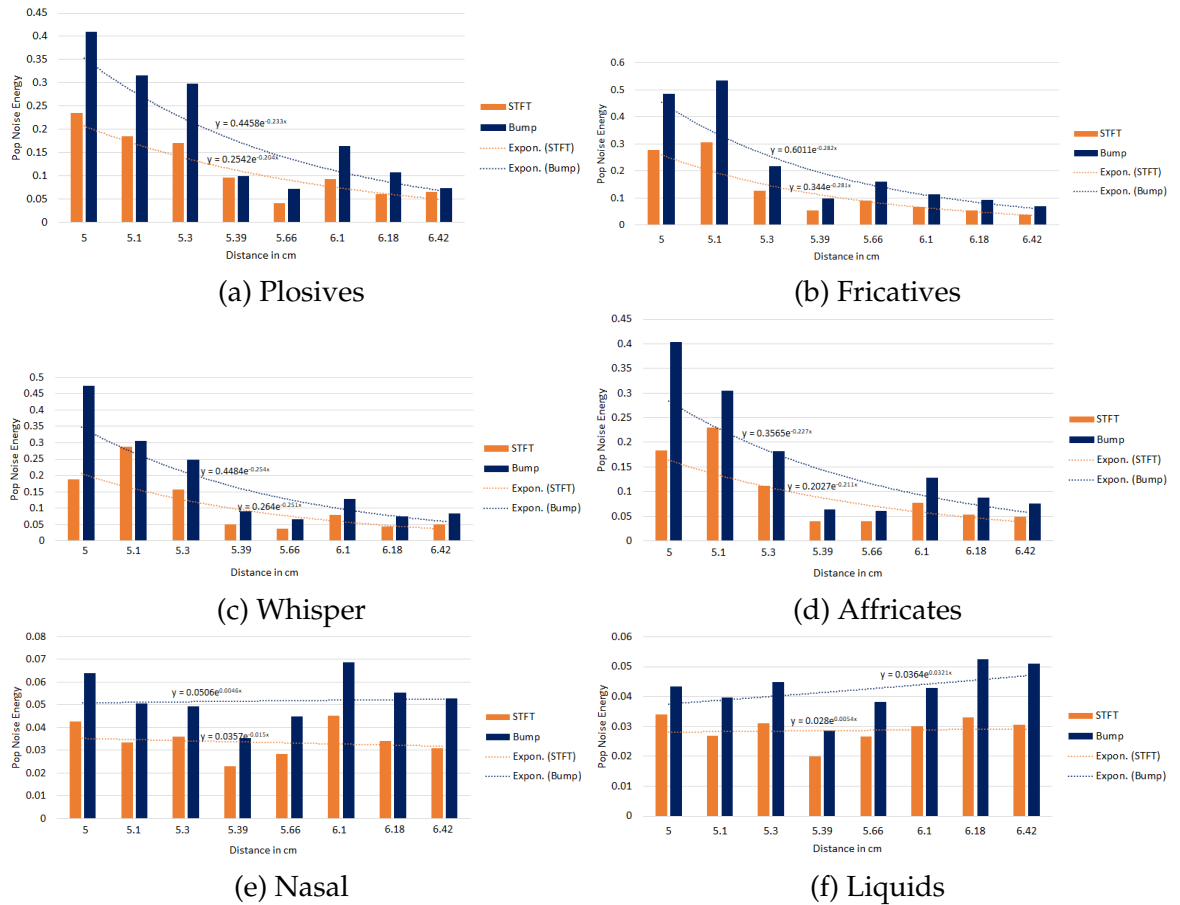


Figure 5.7: Pop noise energies of various phonemes plotted w.r.t. the distance of the speaker from various microphones for the case, when the speaker is at a distance of 5 cm from Mic 7. The pop noise energies are obtained using the proposed Bump wavelet-based approach given via Algorithm 5. The trendlines in each of the sub-figures indicate that the energy of pop noise decreases with the distance of the speaker’s mouth from the microphone.

in volume. For the case of nasal sounds, the spectrum is dominated by the low frequency resonance of the large volume of the nasal cavity [10], thereby having dampened impulse response, which in turn means that the nasal cavity has a large -3 dB bandwidth and hence, relatively more energy loss into the system.

5.3.4 Experimental Results

In this subsection, we present the experimental results for the Bump wavelet-based features using a CNN classifier for the VLD task. We compare the performance of the Bump wavelet-based features with STFT-based baseline features. For fair comparison, both the performances are using CNN as the classifier. We show comparative word-wise VLD accuracies in Figure 5.8. It can be observed from the Figure 5.8 that our proposed bump wavelet-based feature set performs

better for almost every word from the dataset.

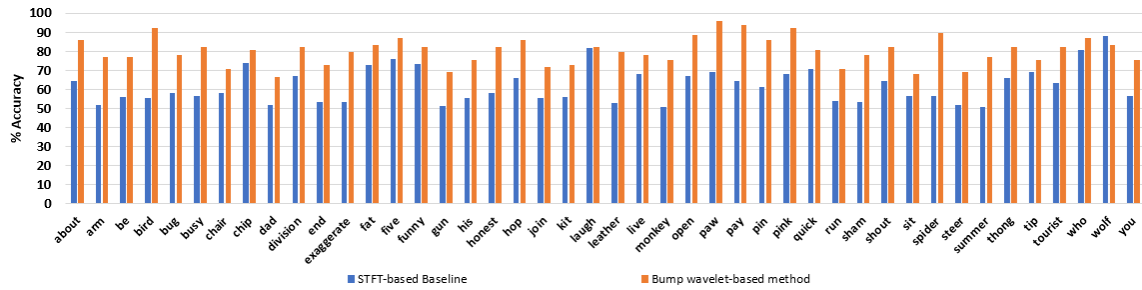


Figure 5.8: Word-wise VLD accuracy of STFT-based baseline method *vs.* proposed bump wavelet-based feature.

Furthermore, we grouped the words in the POCO dataset according to phoneme types as described in Table 3.12 in Chapter 3. To that effect, the Table 5.2 shows the phoneme-wise average VLD accuracy of the proposed method, compared with the existing approaches. We observe that our proposed bump wavelet-based scalogram approach performs better for *all* the phoneme-types. It can be ob-

Table 5.2: Phoneme-wise Average VLD Accuracy (in %).

Phoneme Type	STFT-Based Features (Using SVM) [26]	CQT-Based Features (Using SVM) [102]	STFT-Based Features (Using CNN) [103]	Bump Wavelet-Based Proposed Features (Using CNN) [92]
Plosives	60.46	63.60	71.72	81.58
Whisper	68.44	73.29	76.83	81.09
Fricatives	67.66	73.78	75.55	80.77
Affricates	58.26	68.92	71.83	78.53
Nasal	54.26	57.78	59.33	76.50
Liquids	69.78	57.16	56	69.87

served that using the proposed features, plosive, fricative, and whisper sounds have higher VLD accuracy. This is justified by the nature of pop noise and its dependence on the phoneme type. Plosive, whisper, and fricative sounds have more breathing effects on the microphone as compared to other phonemes [26]. This argument can also be observed from Figure 5.9 and Figure 5.10, where one sample word from each of the phoneme types is taken (i.e., ‘tip’ for plosive, ‘who’ for whisper, ‘laugh’ for fricative, ‘chip’ for affricate, ‘arm’ for nasal, and ‘run’ for liquid phoneme types).

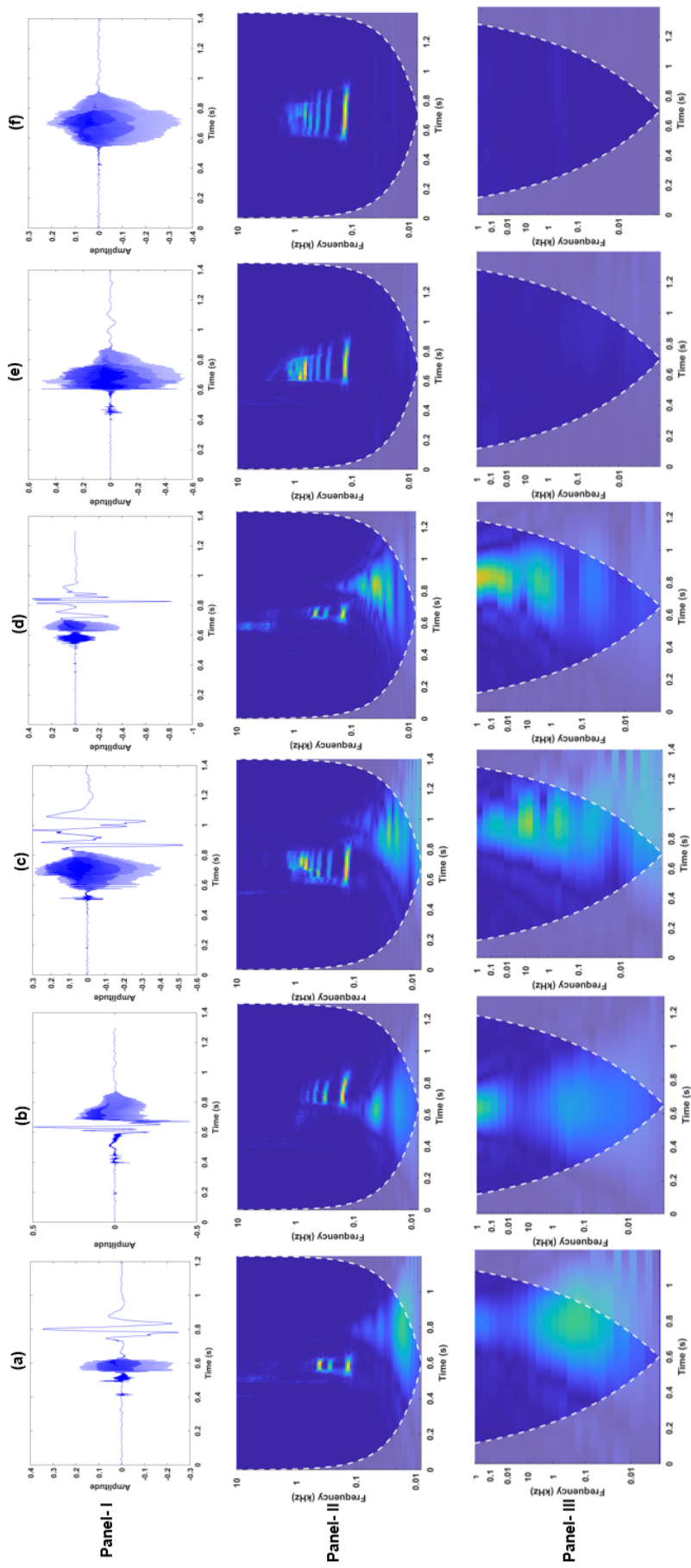


Figure 5.9: Phoneme-wise analysis of genuine (live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Bump wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is present, for (a) plosive (the sample word is 'tip'), (b) whisper (the sample word is 'who'), (c) fricative (the sample word is 'laugh'), (d) affricate (the sample word is 'chip'), (e) nasal (the sample word is 'arm'), and (f) liquid (the sample word is 'run').

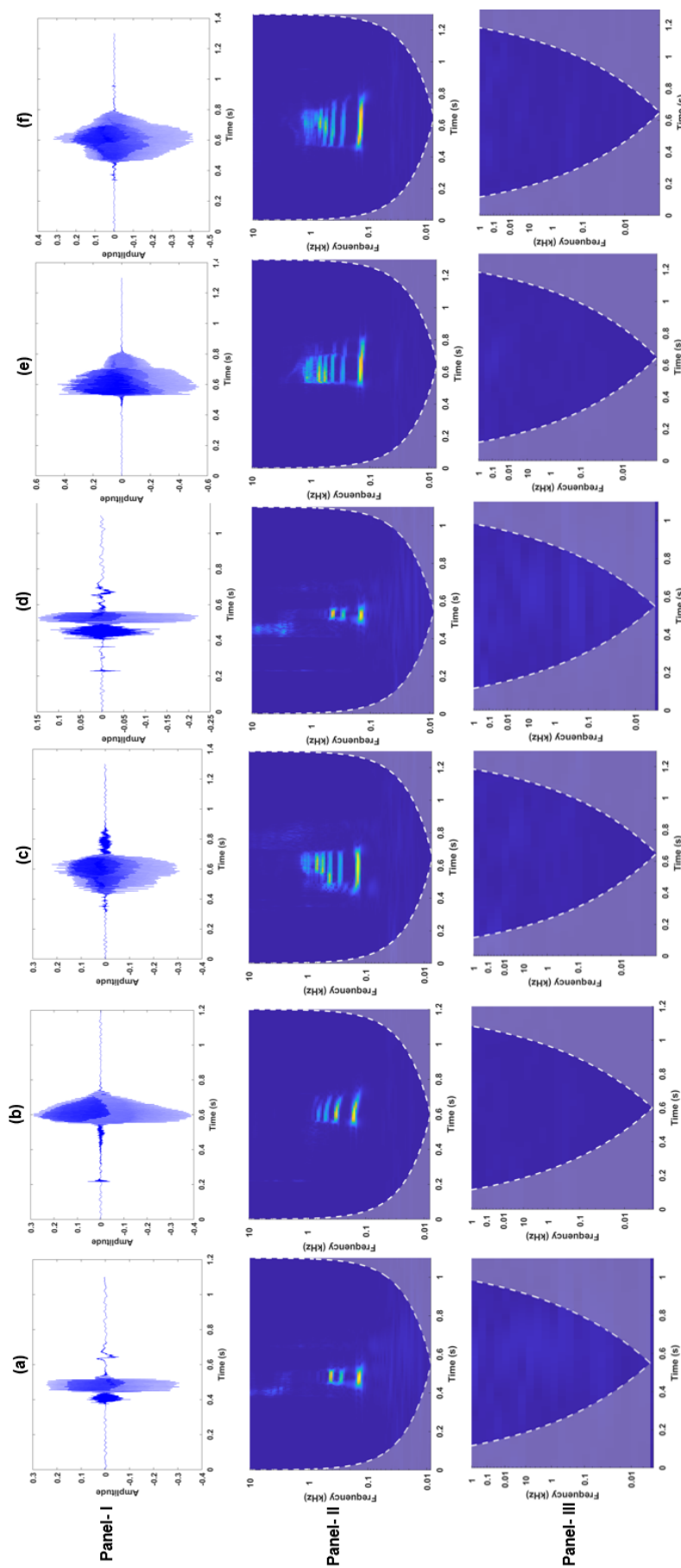


Figure 5.10: Phoneme-wise analysis of spoofed (non-live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Bump wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is absent, for (a) plosive (the sample word is 'tip'), (b) whisper (the sample word is 'who'), (c) fricative (the sample word is 'laugh'), (d) affricate (the sample word is 'chip'), (e) nasal (the sample word is 'arm'), and (f) liquid (the sample word is 'run').

5.4 Morlet Wavelet-Based Features

5.4.1 Proposed Approach

The most famous wavelet w.r.t. the historical development of wavelet research, is the Morlet wavelet, which is a modulated Gaussian, and it is defined as [228]:

$$\psi(t) = e^{j\omega_0 t} e^{-t^2/2}, \quad (5.5)$$

where ω_0 is taken as 5 Hz for a standard Morlet wavelet. The Morlet wavelet is obtained from a Gaussian window multiplied by a sinusoidal wave [6]. We have considered Morlet wavelet because it is closely related to the human perception process (for both hearing and vision) [66]. Moreover, CWT is related to constant-Q filtering- a short-time analysis performed by the peripheral auditory system. In particular, as per original investigations by Flanagan in [229], the wavelet function for the mechanical spectral analysis performed by the Basilar membrane in the cochlea of the human ear is given by $\psi(t) = (t\omega)^2 e^{-t\omega/2}$. Furthermore, Morlet wavelet is the first wavelet (named in honour of its first formal inventor Jean Morlet, even though originally Haar wavelets were formally invented by Haar in 1910 [230]) of its kind in formal historical developments of wavelets in the geophysics literature for the detection of transients and improving the joint time-frequency resolution of seismic signals [231]. Figure 5.11 shows the capturing of pop noise in live (genuine or natural) speech signal using Morlet wavelet-based CWT. It can be observed that the word 'laugh' contains fricative sound (such as, /f/ in 'laugh'), which is produced due to turbulent airflow. It results in bursts of energy at low frequencies for a short-time period, characterizing the presence of pop noise. However, for the case of spoofed speech, the pop noise is not significant, as shown in Panel-II.

5.4.1.1 Handcrafted Morlet Wavelet-Based Features

CWT coefficients are extracted from the speech data of POCO corpus by taking Morlet as the mother wavelet. CWT coefficients are found for frequencies ≤ 40 Hz, as shown in Algorithm 7. Furthermore, to keep the dimension (D) of feature vector as 45 and also to extract the prominent energy of pop noise, the energies are arranged in descending order, and the highest 45-D values are taken for extracting the 45-D feature vector.

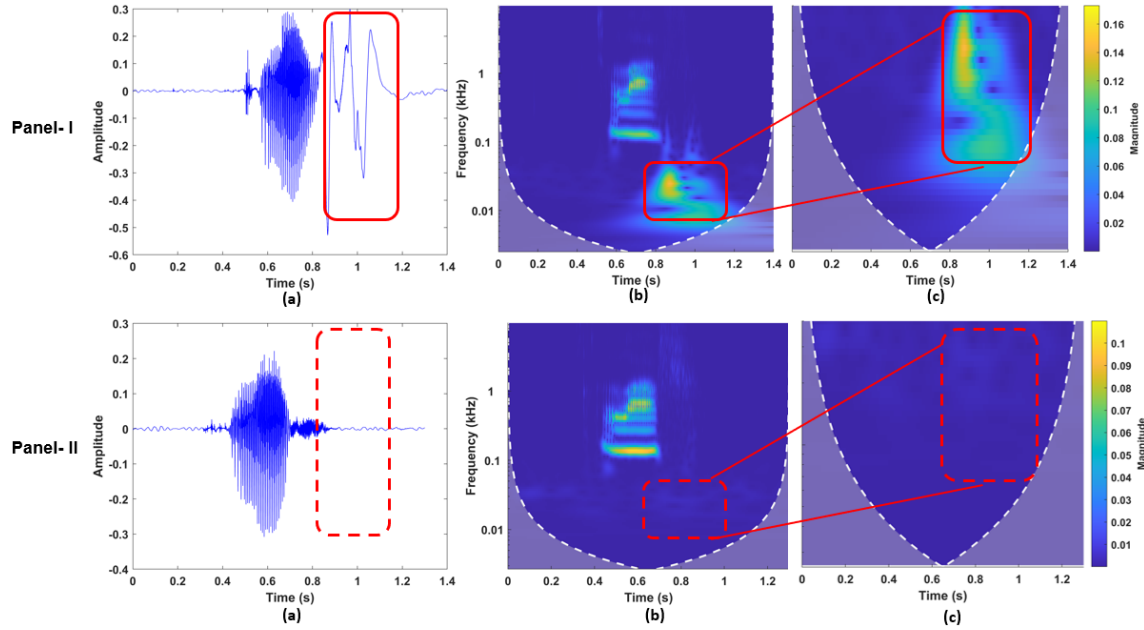


Figure 5.11: Panel I represent the case of presence of pop noise (genuine or live speech). Panel II represents suppressed pop noise (spoofed speech) due to the use of pop filter: (a) time-domain signal for the word 'laugh', (b) corresponding Morlet wavelet-based scalogram, and (c) selected region of scalogram in (b) corresponding low-frequency (0 – 40 Hz). Solid boxes in Panel I indicate the presence of pop noise, while corresponding dotted boxes in Panel II indicates that the pop noise has been suppressed by the use of pop filter.

Algorithm 6 Proposed Handcrafted Morlet Wavelet-based Feature Extraction for VLD.

```

1: procedure FEAT( $x$ )
2:    $w\_name = 'amor'$ 
3:    $[cwt\_coeffs, F] \leftarrow cwt(x, w\_name)$ 
4:    $Low\_bins \leftarrow find(0 < F \leq 40 \text{ Hz})$ 
5:    $Low\_coeffs \leftarrow cwt\_coeffs(Low\_bins)$ 
6:    $Pop\_energy = [abs(Low\_coeffs)]^2$ 
7:    $M = mean(Pop\_energy)$ 
8:    $SD = standard\_deviation(Pop\_energy)$ 
9:    $k \leftarrow length(Low\_coeffs)$ 
10:  while  $r \neq 0$  do
11:     $i = 1$ 
12:     $Norm\_Pop(i) = \frac{Pop\_energy(i) - M}{SD}$ 
13:     $k --, i ++$ 
14:  end while
15:     $[sorted, index] \leftarrow sort(Norm\_Pop, descending)$ 
16:     $Feat \leftarrow Pop\_energy(index(1:dim))$ 
17:  end procedure

```

$\triangleright x$ is the speech signal
 \triangleright Taking Morlet wavelet

5.4.1.2 Low Frequency Morlet Scalogram-Based Features

Since pop noise most likely occurs at frequency regions ≤ 40 Hz, scalograms are very well suited to extract energies at low frequencies because of the higher frequency resolution of scalogram at lower frequencies. For our experiments, the lowest frequency bin is set at 1.9826 Hz. The scale factor between 2 consecutive bins is 1.0718. Therefore, the k^{th} bin index corresponding to 40 Hz is calculated as:

$$40 = (1.0718)^k * 1.9826. \quad (5.6)$$

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

5.4.2 Setup

Similar to the setup as described in subsection 5.3.2, the POCO dataset is used for the experiments on Morlet wavelet-based features. However, the data is partitioned into training, Dev, and Eval sets, as described in Table 3.11 in Chapter 3. CNN is used as the classifier with the same architecture as used for Bump wavelet-based features in subsection 5.3.2. Similarly, for speaker-microphone distance-wise analysis, RC-B subset of the POCO dataset is used with the microphone arrangement and distance calculation as described in Table 3.10 in Chapter 3.

5.4.3 Speaker-Microphone Distance-Based Analysis

In this subsection, we show the effect of distance variability on the strength of pop noise. To that effect, Figure 5.12 shows 3 cases, where the distance between a speaker's mouth and the microphone is varied as 5 cm, 10.78 cm, and 20.48 cm, as Panel-I, II, and III, respectively. The word spoken is '*pink*' taken from the POCO dataset. It can be observed that in time-domain representation of the signal, the pop noise is dominantly visible in Panel-I, where the speaker's mouth is the closest to the microphone (i.e., 5 cm). Similar observation can be made from its CWT-based full-frequency as well as low-frequency scalogram representations, where pop noise energy is highlighted in red boxes. Next, Panel-II shows when the speaker's mouth is at a distance of 10.78 cm from the microphone. It can be

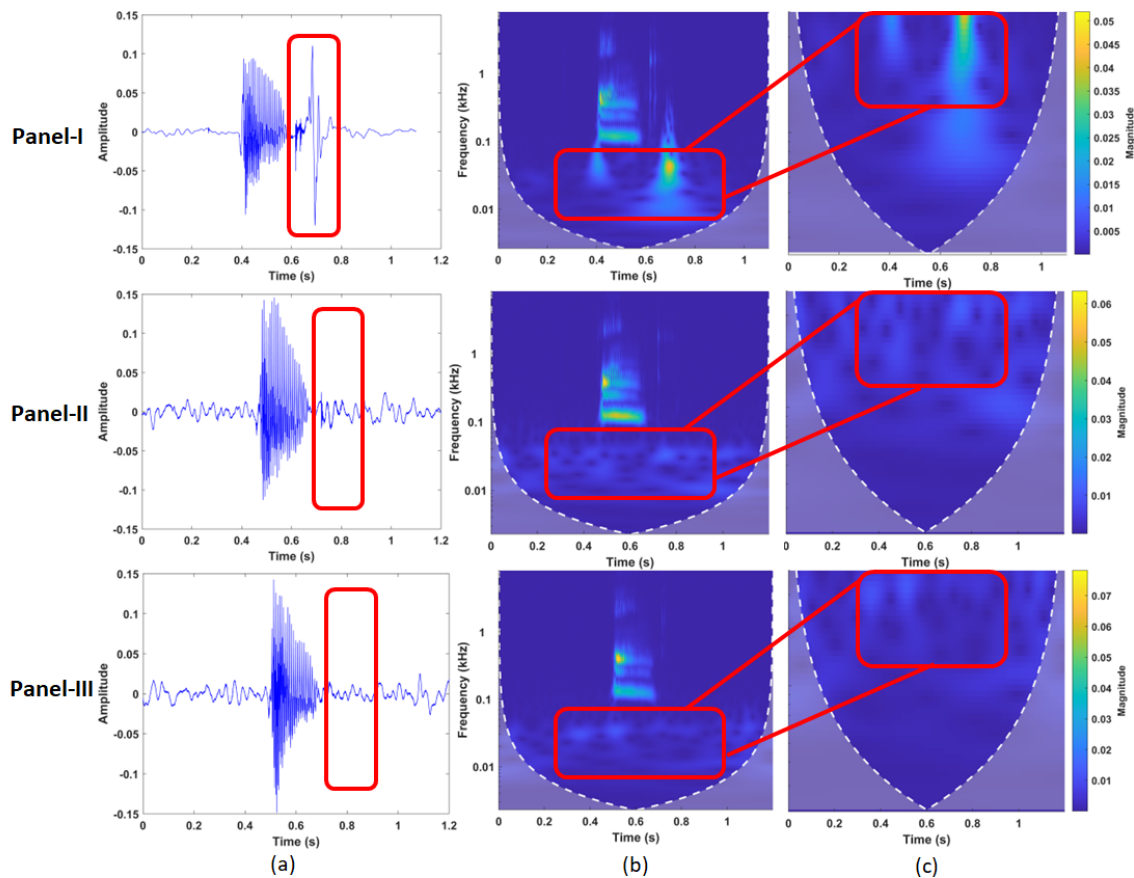


Figure 5.12: Panel I, Panel II, and Panel-III represent the varying distance of the speaker from the microphone, namely, 5 cm, 10.78 cm, and 20.40 cm, respectively, for (a) time-domain signal for the word ‘pink’, (b) corresponding scalogram, and (c) selected region of scalogram corresponding low-frequency (0 – 40 Hz). Solid boxes in red indicate the presence of pop noise.

observed from Panel-II that the pop noise is not visible dominantly in the time-domain representation. On the other hand, the scalogram-based time-frequency representation is able to capture the pop noise energy in the low frequency regions. However, the strength of pop noise energy is degraded as compared to Panel-I. Lastly, Panel-III shows the case, when the speaker’s mouth is at the farthest distance from the microphone, i.e., 20.40 cm. One can observe the lowest strength of pop noise energy in this case, for time-domain representation, as well as the scalogram-based representations. For analysis purposes, in this subsection, we considered a particular word ‘pink’ as an example to show the effect of distance. However, it should be noted that the word ‘pink’ contains plosives predominantly. Apart from the distance, the strength of pop noise captured also depends on the type of phonemes present in the word uttered in front of the microphone.

Figure 5.13 shows the various cases of phoneme categories and the effect of

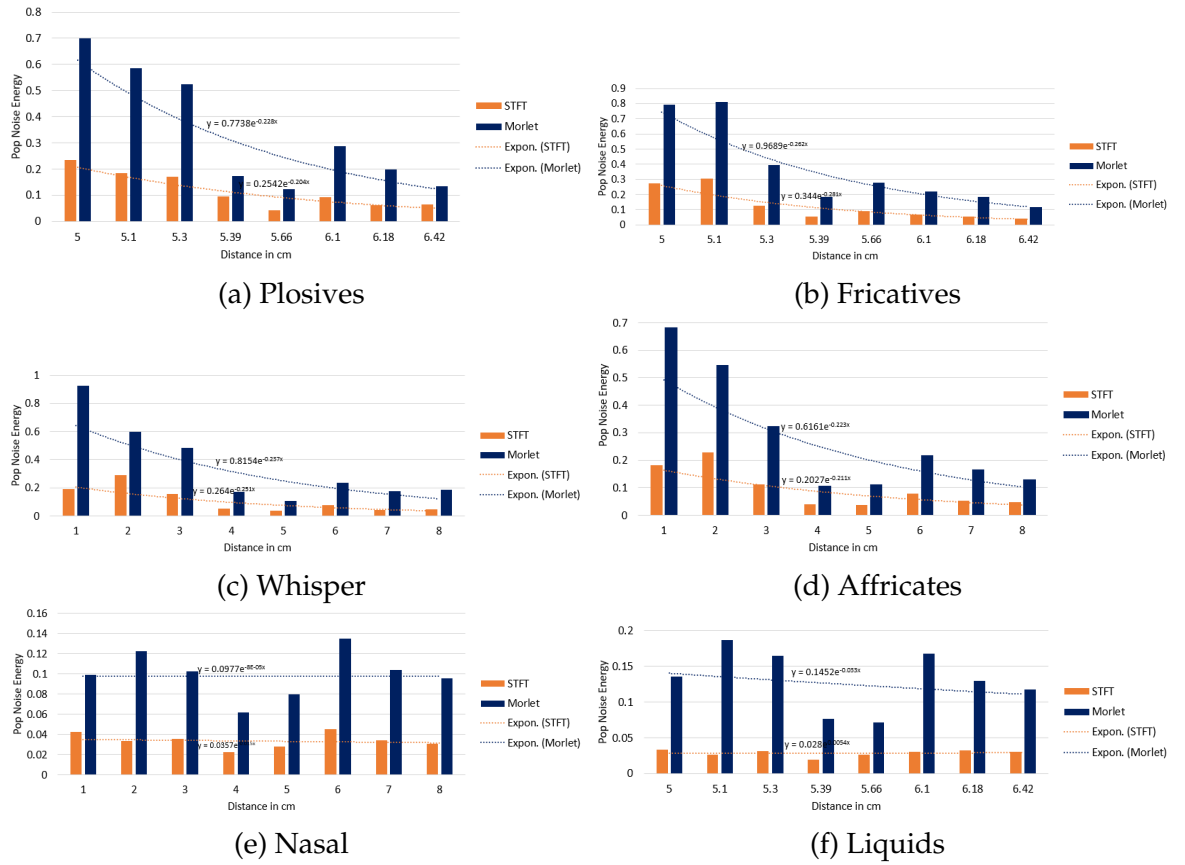


Figure 5.13: Pop noise energies for various phoneme sounds plotted w.r.t. the distance of the speaker from various microphones for the case when the speaker is at a distance of 5 cm from Mic 7. The pop noise energies are obtained using the proposed Morlet wavelet-based Algorithm 7. The dotted curve in each of the sub-figures indicates that the energy of pop noise decreases with the distance of the speaker's mouth from the microphone.

Algorithm 7 Proposed Algorithm for Pop Noise Energy Estimation Using Bump Wavelet for VLD.

```

1: procedure ENERGY_POP( $x$ )
2:    $w\_name = 'amor'$ 
3:    $[cwt\_coeffs, F] \leftarrow cwt(x, w\_name)$ 
4:    $Low\_bins \leftarrow find(0 < F \leq 40 \text{ Hz})$ 
5:    $Low\_coeffs \leftarrow cwt\_coeffs(Low\_bins)$ 
6:    $Pop\_energy = [abs(Low\_coeffs)]^2$ 
7:    $[r, c] = size(Pop\_energy)$ 
8:   for  $i \leftarrow 0 : r$  do
9:      $E\_LF(i) = sum(Pop\_energy(i, :))$ 
10:  end for
11:   $E_{mean} = mean(E\_LF)$ 
12: end procedure

```

$\triangleright x$ is the speech signal
 \triangleright Taking Morlet wavelet
 \triangleright Each row of E_LF has energy for 1 frequency bin

distance on the strength of the pop noise. For this analysis, the RC-B subset of the POCO dataset is used in this work. The trend of pop noise energy is shown with the help of the dotted line. In particular, we observe a somewhat similar trend for four of the phoneme categories, namely, plosives, fricatives, whisper, and affricates. However, for nasal and liquids, we see almost constant-like trendlines (with equations as $y = 0.0977e^{-10^{-5}x}$ and $y = 0.1452e^{-0.033x}$, respectively), which shows that the distance does not affect the pop noise energies in nasals and liquids categories of phonemes.

Furthermore, the existing results (in the form of % classification accuracy for the VLD task) in the literature as shown in Table 5.3, also show that the best performance is achieved using the Morlet wavelet-based scalogram approach. In particular, for *all* the phoneme classes, the proposed system shows relatively the best performance. Furthermore, it should also be noted that the lowest VLD accuracies of 80.77% and 79.49% are obtained on nasal and liquid sounds. A similar observation can be made by the analysis done in Figure 5.13, wherein the nasal and liquids have the *least* pop noise energies (as shown in Figure 5.13 (e) and (f)). In addition to this, the trendlines show that these two classes of phonemes are the *least* affected by the distance of the microphone from the speaker. Thus, our results and analysis presented via Figure 5.13 are in strong agreement with the recent results reported on VLD task in [28].

5.4.4 Experimental Results

5.4.4.1 Proposed Handcrafted Morlet-Based Features

For the case of 45-D wavelet-based features (indicated as system (F)), we achieved an overall VLD accuracy of 80%. Figure 5.16 shows word-wise VLD accuracy over 44 words in the dataset. We observed that the word ‘pay’ has the highest accuracy of 91.02%, because the word ‘pay’ has a strong plosive sound of /p/. Furthermore, we achieved an average accuracy of 79.35% and 79.27% on words with prominent performance on plosives and fricatives, respectively, as shown in Table 5.3. Correspondingly, one example from each phoneme type is taken ((i.e., ‘tip’ for plosive, ‘who’ for whisper, ‘laugh’ for fricative, ‘chip’ for affricate, ‘arm’ for nasal, and ‘run’ for liquid phoneme types)) and the corresponding Morlet wavelet-based scalograms are analysed in Figure 5.14 and Figure 5.15, for genuine and spoofed replay cases, respectively.

5.4.4.2 Proposed Morlet Scalogram-Based Features

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

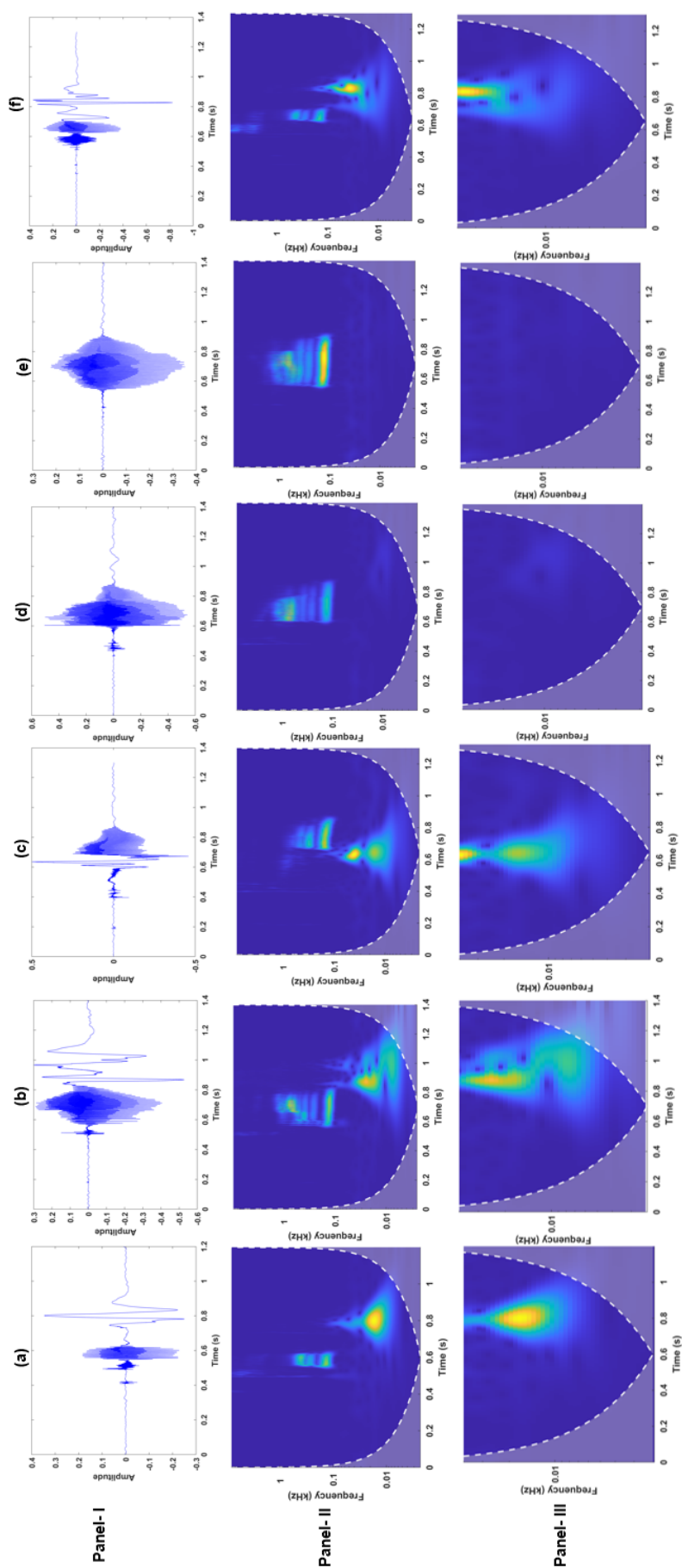


Figure 5.14: Phoneme-wise analysis of genuine (live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morlet wavelet-based scalogram where pop noise is present, for (a) plosive (the sample word is 'tip'), (b) fricative (the sample word is 'laugh'), (c) whisper (the sample word is 'who'), (d) nasal (the sample word is 'arm'), (e) liquid (the sample word is 'run'), and (f) affricate (the sample word is 'chip').

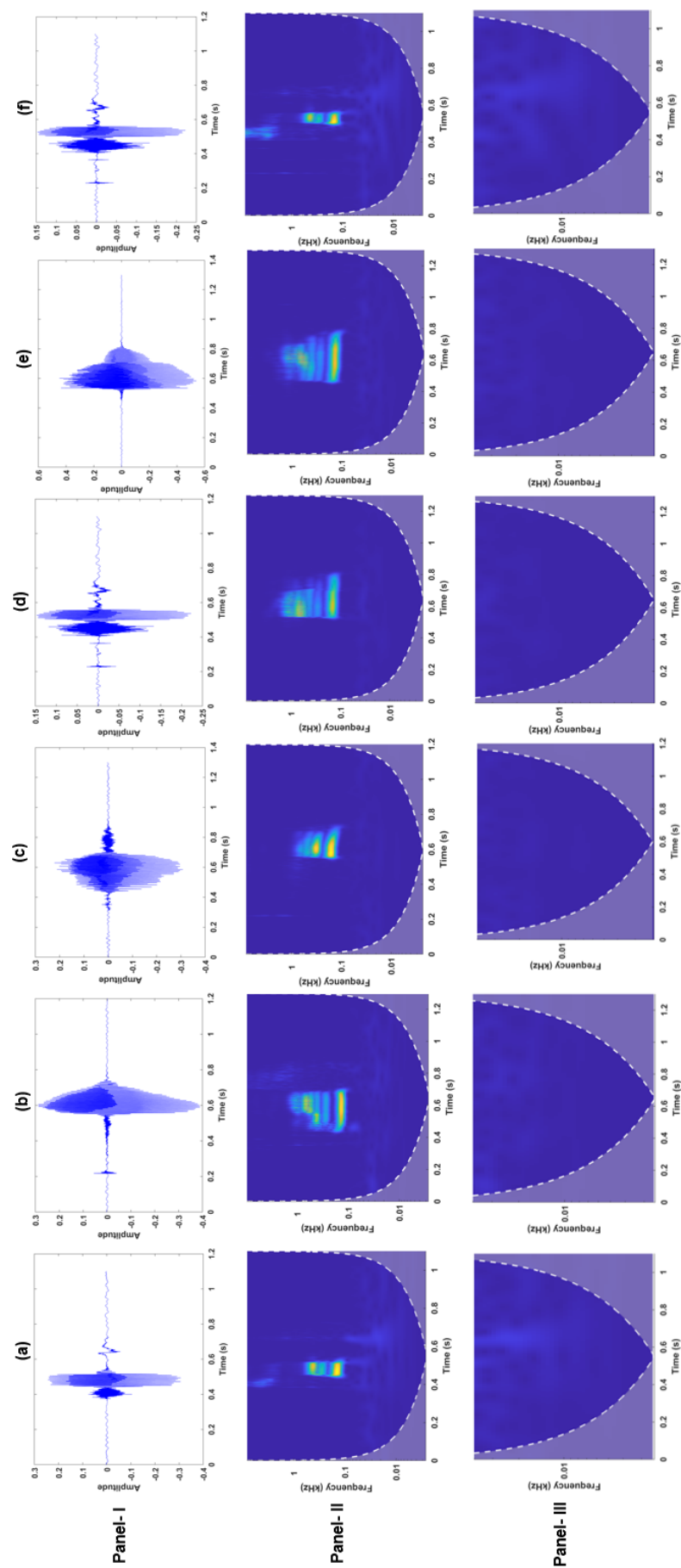


Figure 5.15: Phoneme-wise analysis of spoofed (non-live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morlet wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is absent, for (a) plosive (the sample word is 'tip'), (b) fricative (the sample word is 'laugh'), (c) whisper (the sample word is 'who'), (d) nasal (the sample word is 'arm'), (e) liquid (the sample word is 'run'), and (f) affricate (the sample word is 'chip').

Table 5.3: Average VLD Accuracy (in %) of Different Phoneme Types.

Phoneme Type	(A) Spectrogram (SVM) [26]	(B) CQT (SVM) [102]	(C) Spectrogram (CNN) [103]	(D) Mel-spectrogram (CNN)	(E) Handcrafted Bump Wavelet-based (CNN) [92]	(F) Handcrafted Morlet Wavelet-based (CNN) (Proposed)	(G) Handcrafted Morlet Scalogram (CNN) (Proposed)
Freq. Range	0-40 Hz	0-11025 Hz	0-11025 Hz	0-40 Hz	0-40 Hz	0-40 Hz	0-40 Hz
Plosive	60.46	63.61	71.72	74.13	81.58	79.35	89.07
Fricatives	67.66	73.78	75.55	77.45	80.77	79.27	87.61
Whisper	68.44	73.29	76.83	74.99	81.09	79.48	86.21
Nasal	54.26	57.78	59.33	70.51	76.50	71.36	80.77
Liquids	69.78	57.16	56	69.23	69.87	65.38	79.49
Affricates	58.26	68.92	71.83	72.51	78.53	74.35	85.26

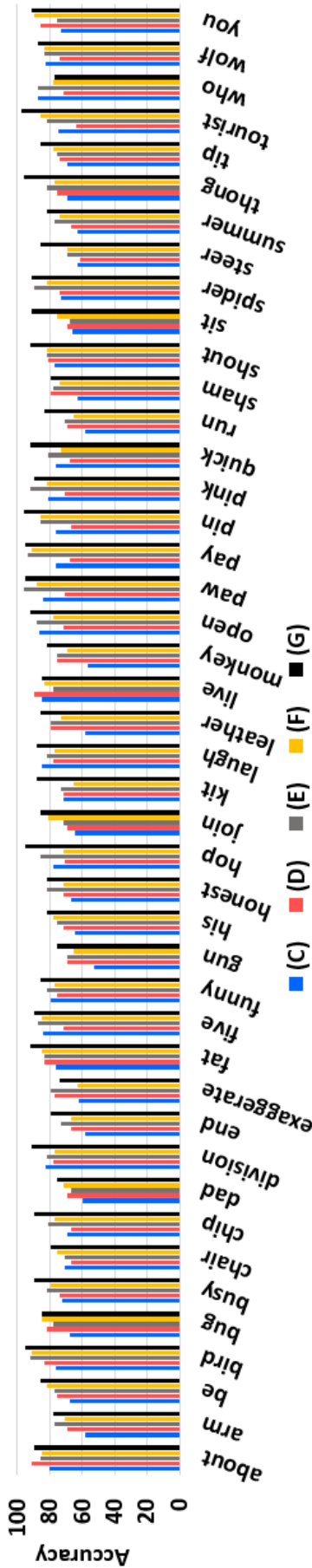


Figure 5.16: Word-wise VLD accuracies (in %) with CNN classifier for (C): Full frequency spectrogram, (D): Low frequency Mel-spectrogram, (E): Handcrafted Bump wavelet-based features, (F): Handcrafted Morlet wavelet-based features, and (G): Handcrafted Morlet scalogram. The indices (C)-(G) are w.r.t. the labels in Table 5.3.

It can be observed from Table 5.3 that the proposed Morlet scalogram-based approach outperforms every other method for *all* the phoneme types. Furthermore, all the methods are found to perform relatively better for plosives and fricative sounds. Fricative sounds (such as, /f/sound in the word ‘laugh’) are produced due to turbulent airflow, which results in bursts of energy at low frequencies for a short-time period, characterizing the presence of pop noise. Furthermore, plosive sounds (such as, /p /sound in ‘pay’) are caused by a sudden release of a burst of air from the lips, resulting in pop noise [10]. On the contrary, energy distribution in nasal sounds is due to air flow in the nasal cavity, while the oral cavity is closed, and therefore the sound is radiated at the nostrils [10].

5.5 Generalized Morse Wavelet (GMW)-Based Features

Given that there exists an issue of selecting an appropriate wavelet, we now proposed to exploit Generalized Morse Wavelet (GMW)-based features for the VLD task. GMWs act as a superfamily of analytic wavelets. Moreover, they show strictly analytic properties, which is defied by Morlet wavelet under some conditions. The details of the motivation and advantages of GMWs are discussed in detail in subsection 5.5.1.2.

5.5.1 Proposed Approach

5.5.1.1 Generalized Morse Wavelets (GMWs)

GMWs are defined in the frequency-domain as [9,232]:

$$\Psi_{\beta,\gamma}(\omega) = \int_{-\infty}^{\infty} \psi_{\beta,\gamma}(t)e^{-i\omega t} dt = U(\omega)a_{\beta,\gamma}\omega^{\beta}e^{-\omega^{\gamma}}, \quad (5.7)$$

where $U(\omega)$ is the unit-step function in the frequency-domain, and $a_{\beta,\gamma}$ is the normalizing constant such that

$$a_{\beta,\gamma} \equiv 2 \left(\frac{e\gamma}{\beta} \right)^{\beta/\gamma}, \quad (5.8)$$

where e is Napier’s number, which is commonly defined as the base of the natural logarithm. Furthermore, the parameters β and γ provide an additional degree of freedom, and make GMWs form a *family* of analytic wavelets [9]. For $\gamma = 1$, these wavelets become equivalent to a solution to the Schrödinger equation examined by Morse in [233].

The behaviour of GMWs can be characterized using normalized versions of the derivatives of the frequency-domain representation of the wavelet [232]. Therefore, the wavelet's *dimensionless derivatives* are defined as [9,232]:

$$\tilde{\Psi}_n(\omega) \equiv \omega^n \frac{\Psi^{(n)}(\omega)}{\Psi(\omega)}, \quad (5.9)$$

where n as superscript denotes the n^{th} order derivative. To that effect, for analysis purposes, a conventional dimensionless parameter based on the 2^{nd} order derivative is defined as $P_{\beta,\gamma}$, which is mathematically expressed as [9]:

$$P_{\beta,\gamma} \equiv \sqrt{-\tilde{\Psi}_{2;\beta,\gamma}(\omega_\psi)} = \sqrt{-\omega_\psi^2 \frac{\Psi''(\omega_\psi)}{\Psi(\omega_\psi)}} = \sqrt{\beta\gamma}, \quad (5.10)$$

where $\omega_\psi \equiv (\beta/\gamma)^{1/\gamma}$ is the peak frequency at which the frequency-domain wavelets obtain a maximum value. In the eq. (5.10), the 2^{nd} order derivative of $\Psi(\omega)$ (i.e., $\Psi''(\omega_\psi)$) is computed at the peak frequency, ω_ψ . At the peak frequency, there is a maxima of $\Psi(\omega)$. Therefore, $\Psi''(\omega_\psi)$ is negative. Given it is an analytic wavelet, $\Psi(\omega_\psi)$ will be positive always and hence, the complete term inside the underroot will be positive. Therefore, the parameter $P_{\beta,\gamma}$ is real-valued, and it measures the *duration* of the wavelet function. To that effect, $P_{\beta,\gamma}/\pi$ mea-

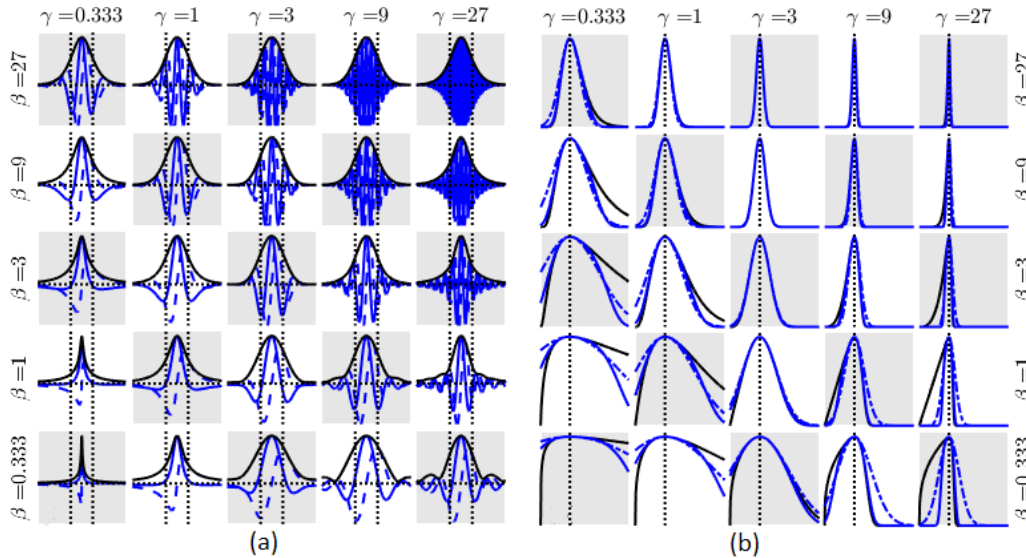


Figure 5.17: Morse wavelets for varying values of β and γ in (a) time-domain, and (b) frequency-domain. After [8].

asures the number of oscillations at the peak frequency, which fit within the central time window of the wavelet function [9]. Furthermore, $1/P_{\beta,\gamma}$, which is the inverse of the time domain duration, can be seen to be a measure of the *bandwidth* of

the wavelet. In other words, increasing the value of $P_{\beta,\gamma}$ increases the frequency-domain curvature in the vicinity of the peak frequency, therefore, narrowing the wavelet in the frequency-domain, and dilating the wavelet in the time-domain [9]. Furthermore, the parameter pertaining to the 3^{rd} order dimensionless derivative at the peak frequency is [9]:

$$\tilde{\Psi}_{3;\beta,\gamma}(\omega_\psi) = -(\gamma - 3)P_{\beta,\gamma}^2. \quad (5.11)$$

Since the Morse wavelets are parameterized by two parameters β and γ , their 2^{nd} and 3^{rd} order properties $P_{\beta,\gamma}$ and $\tilde{\Psi}_{3;\beta,\gamma}(\omega_\psi)$ can be varied independently to generate popular families of analytic wavelets. Figure 5.17 shows the effect of β and γ individually, on the shape of the Morse wavelet in time as well as frequency-domains.

The GMWs superfamily is said to unify almost all the analytic wavelet families, such as Morlet, Derivative of Gaussian, Cauchy-Klauder-Morse-Paul, log-normal, Bessel, and Shannon wavelets. As shown in Figure 5.18, for the Morse

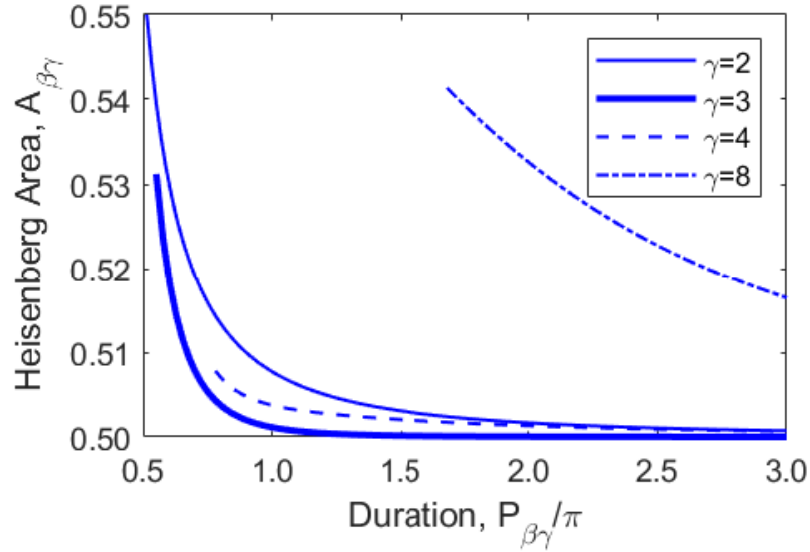


Figure 5.18: Effect of γ parameter on the time-frequency Heisenberg area $A_{\beta,\gamma}$ w.r.t. wavelet duration $P_{\beta,\gamma}/\pi$. After [8].

wavelet with parameter $\gamma = 3$ (also known as '*Airy family*'), we get the most optimum Heisenberg area $A_{\beta,\gamma}$ for $\gamma = 3$ even for a small wavelet duration [8]. The area of Heisenberg's box $A_{\beta,\gamma}$ is [232]:

$$A_{\beta,\gamma} = \sigma_t^2 \sigma_\omega^2, \quad (5.12)$$

where σ_t^2 and σ_ω^2 are the *time* and *frequency* spread of the wavelet atom, respec-

tively. The minimum value of $A_{\beta,\gamma}$ is governed by $\sigma_t^2 \cdot \sigma_\omega^2 \geq 1/4$, where σ_t^2 and σ_ω^2 are the variances in time and frequency-domains, respectively [6].

The name 'Airy family' for $\gamma = 3$ comes from the existence of a 2nd order linear differential equation known as the 'Airy' equation or the Stoke's equation, given by [234]:

$$\frac{d^2y}{dx^2} - xy = 0, \quad (5.13)$$

whose solution changes from oscillatory to exponential functions. A special type of 2nd order differential equation is given by

$$\frac{d^2y}{dx^2} - xy = \frac{1}{\pi}, \quad (5.14)$$

whose solutions are called as *Scorer's functions* and are denoted as $H_i(x)$ and $G_i(x)$. The Scorer's functions are given by [234]:

$$G_i(x) = \frac{1}{\pi} \int_0^\infty \sin\left(\frac{t^3}{3} + xt\right) dt, \quad (5.15)$$

$$H_i(x) = \frac{1}{\pi} \int_0^\infty \exp\left(-\frac{t^3}{3} + xt\right) dt. \quad (5.16)$$

The functions $H_i(x)$ and $G_i(x)$ are called as 1st and 2nd Scorer functions, respectively. The GMW corresponding to $\gamma = 3$ comes from $H_i(z)$ (i.e., by replacing x with z in eq. (5.16)), where

$$H_i(z) = \frac{1}{\pi} \int_0^\infty e^{(-u^3/3)} e^{zu} du. \quad (5.17)$$

GMW in time-domain is $IFT\{\Psi(\omega)\} = \frac{1}{2\pi} \int_0^\infty \Psi(\omega) e^{j\omega t} d\omega$, which can be expressed as

$$\psi_{\beta,\gamma}(t) = \frac{1}{2\pi} \int_0^\infty a_{\beta,\gamma} \omega^\beta e^{-\omega^\gamma} e^{j\omega t} d\omega. \quad (5.18)$$

For $\beta = 0$, and $\gamma = 3$, eq. (5.18) reduces to

$$\psi_{0,3}(t) = \frac{1}{\pi} \int_0^\infty e^{-\omega^3} e^{j\omega t} d\omega, \quad (5.19)$$

which is nothing but $\frac{1}{3^{1/3}} H_i\left(\frac{jt}{3^{1/3}}\right)$, i.e., the time-domain representation of Morse wavelet for $\beta = 0$ and $\gamma = 3$.

5.5.1.2 Advantage of GMW

As discussed in the previous subsection, GMWs take the form of various types of analytic wavelets, depending on the value of the parameters of the Morse wavelet. In particular, GMWs at $\gamma = 3$ take the form of Morlet wavelet.

Morlet wavelet is a sinusoid modulated by a Gaussian function. Though it has infinite duration, most of its energy is confined to a finite interval, and it is suited for good time and frequency-localized information because of its equal variance in time and frequency-domains. However, the mother wavelet $\psi(t)$ does not satisfy the admissibility condition *exactly* due to its infinite duration. For $\omega_0 \geq 5$, the error due to violation of the admissibility condition can be ignored. The Morlet wavelet is defined as [6,228]:

$$\psi(t) = e^{j\omega_0 t} e^{-t^2/2}, \quad (5.20)$$

where ω_0 is the frequency of the sinusoid in the mother wavelet, and its Fourier transform is given by [6]:

$$\begin{aligned} \Psi(\omega) &= \mathbb{F}\{\psi(t)\} = \mathbb{F}\{e^{j\omega_0 t - t^2/2}\}, \\ &= \pi^{3/2} e^{-(\omega - \omega_0)^2/4}. \end{aligned} \quad (5.21)$$

Thus, the analyticity of Morlet wavelet is controlled by the choice of ω_0 . However, despite such significance of Morlet wavelet, we choose to consider GMWs at $\gamma = 3$, instead of the conventional Morlet wavelet. The motivation to do so is validated by the analysis shown in this subsection. In particular, Figure 5.19 shows the comparison of Morse (at $\gamma = 3$), and the conventional Morlet wavelet. For simplicity, we will now refer to the conventional Morlet wavelet as the Morlet wavelet, and the Morse wavelet at $\gamma = 3$ as the Morse wavelet. On comparison of Morlet *vs.* Morse, we observe local minima in the modulus of the Morlet wavelet as shown in Figure 5.19 (a). Its corresponding Wigner-Ville distribution is shown in Figure 5.19 (e). Wigner-Ville distribution is a time-frequency representation with minimum loss of resolution. Unlike the time-frequency representations, such as spectrograms and scalograms (which are computed by correlating the signal under consideration with families of time-frequency atoms), Wigner-Ville distribution is computed by correlating the signal under consideration with a time-frequency translation of itself. In other words, the time and frequency resolution of spectrograms and scalograms is limited by the time-frequency resolution of the corresponding atoms. Therefore, for analysis purposes Wigner-Ville distribution is estimated. However, it suffers from the issue of the existence of interference

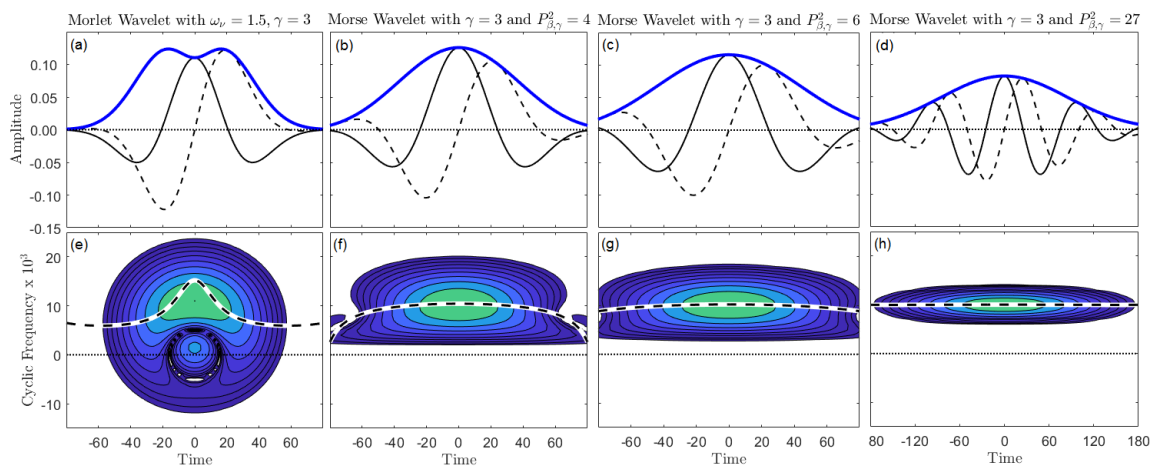


Figure 5.19: Illustration of spectral leakage in (a) Morlet wavelet from 'Airy' family, vs. (b-d) Morse wavelets with $\gamma = 3$, and varying $P_{\beta,\gamma}^2$ values, and their respective Wigner-Ville distributions shown in (e)-(h). After [9].

terms, which makes the applications of Wigner-Ville distribution to be limited [6].

Figure 5.19 (e) shows that there is *spectral leakage* in the negative frequency region. This issue of spectral leakage in negative frequencies is obliterated in the case of Morse wavelets. For fair comparison with Morlet wavelet, the value of γ of Morse wavelet is kept fixed as 3. To that effect, variations of Morse wavelet are shown in Figure 5.19 (b)-(d). Their corresponding Wigner-Ville distributions are shown in Figure 5.19 (f)-(h). It can be observed that for *all* the variants of Morse wavelets, there is *no* spectral leakage in the negative frequency regions. Therefore, we can say that the Morse wavelet exhibits *strict* analytic properties (i.e., $\Psi(\omega) = 0$ for $\omega < 0$), whereas the Morlet wavelet is not strictly analytic (i.e., $\Psi(\omega) \neq 0$ for $\omega < 0$). The authors believe that this strict analytic characteristic of GMWs helps to effectively capture pop noise events that are likely to be present near 0 to 40 Hz, whereas for analytic wavelets such as the Morlet wavelet, the energy in this low-frequency region may experience a blur due to spectral leakage in the negative frequencies, and also the fixed amount of signal energy that is represented in the time-frequency with Parseval's energy equivalence [235].

5.5.1.3 Morse Wavelet-Based Features for the VLD task

As discussed in the previous subsection, Morse wavelet is predominantly parameterized by two entities- γ and $P_{\beta,\gamma}^2$. In the previous subsection, we observed that at $\gamma = 3$, Morse wavelet is in close approximation with the Morlet wavelet, which is known to capture perceptual cues effectively (both in visual and hearing domains). Furthermore, Morlet wavelet is observed to defy analytic behaviour, un-

like Morse wavelet at $\gamma = 3$, which is strictly analytic. To that effect, Figure 5.20 shows the analysis of pop noise using Morse wavelet-based scalogram for $\gamma = 3$. Panel-I denotes the case of genuine (live) speech, and Panel-II denotes the case of

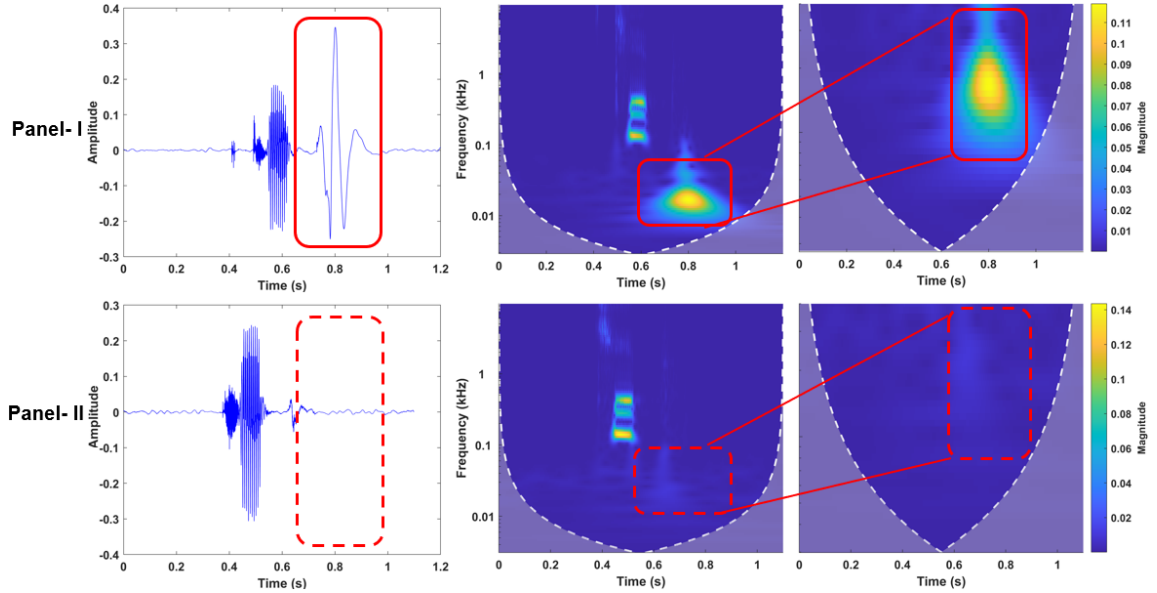


Figure 5.20: Panel I: Genuine speech *vs.* Panel II: Spoofed speech, (a) time-domain signal for the word ‘tip’, (b) corresponding Morse wavelet-based scalogram, and (c) corresponding low frequency (0 – 40 Hz) scalogram.

spoofed speech. It can be observed that due to the improved frequency resolution in the lower frequency regions, the Morse wavelet-based scalogram has the capability to capture pop noise effectively in genuine speech signal. In the literature, regions below 40 Hz are known to have pop noise predominantly and hence, for the analyses presented in this subsection, we have considered the low frequency region to be < 40 Hz as the initial setting for the experiments in this work.

5.5.2 Setup

- Dataset Used:** For the experiments, POCO dataset is used, wherein RC-A and RP-A subsets are used for classification between live and non-live utterances (details given in subsection 3.3.1 of Chapter 3). In addition, to observe the effect of room acoustics, the impulse response of a simulated reverberation environment is created. This is done by generating two-channel impulse response w.r.t. room conditions, as in the ASVSpooF 2019 challenge dataset [236–238], with parameters and configurations as shown in Table 5.4. The source position is varied from 10 cm to 90 cm for observing the effect of variation in distance between the speaker and attacker’s recording

device. This impulse response is then convoluted with the utterances of the POCO dataset, to give us a new subset of the dataset, namely REP-A, as done in [93].

Table 5.4: Parameters and the corresponding configurations for replay mechanism. After [30].

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

- **Classifiers used:**

- **CNN:** In this work, CNN architecture having 3 convolutional layers is used, with each layer having kernel sizes of 7×7 , 3×3 , and 3×3 . The number of input channels of the three convolutional layers are 3, 16, and 32, with the number of output channels of the final layer as 64. Each convolutional layer is followed by batch normalization, and ReLU is used as the activation function. Furthermore, 3 FC layers are used, with max-pooling operation having kernel size of 3×3 , and a stride of 3. The learning rate is kept as 0.001, with BCE as the loss function, and Adam optimizer for optimization of the weights.
- **LCNN:** Another neural network-based classifier used in this work is LCNN, as it is also one of the architectures used for replay SSD [158]. It uses a special case of max-out known as Max Feature Map (MFM) activation function, which is defined as [239]:

$$y_{ij}^k = \max \left(x_{ij}^k, x_{ij}^{k+\frac{N}{2}} \right), \quad (5.22)$$

where the number of channels in the first convolutional layer is $2N$, where $1 \leq k \leq N$, $1 < j \leq W$, and $1 \leq i \leq H$. Here, i and j represent the feature component and frame number, respectively.

For our experiments, the LCNN model consists of three CNN layers (Conv1, Conv2, and Conv3) and one FC layer (FC1). The weights are initialized using Xavier’s normalization method. The data are convoluted using a kernel of size 3×3 and a stride of 1 in the convolutional layers. The MFM and max-pooling layer are applied after each

layer. The MFM layer employs a 3×3 size kernel with a stride of 1 and padding of 2. The model's complexity is decreased by using the max-pooling with a 2×2 kernel size and stride of 2. In the FC1 layer, the ReLU activation function is employed to distinguish between genuine and spoofed classes. We used the BCE loss function to calculate the loss, and the stochastic gradient descent approach to optimize the weights.

- **ResNet:** In this work, the ResNet architecture consists of 1 convolutional layer, 4 residual blocks, and 1 fully-connected layer. The convolutional layer has kernel size of 7×7 . Furthermore, each residual block consists of two blocks of two convolutional layers each, followed by ReLU as the activation function, where each convolutional layer has kernel size of 3×3 .
- **RawNet2:** In this study, the RawNet2 architecture is used because it has recently been proposed and is a successful architecture for the SSD problem [240]. It was used as one of the baseline architectures in the ASVSpooof-2021 challenge. The RawNet2 developed from RawNet, an end-to-end deep neural network that was proposed for text-independent speaker verification [240]. Following the Gated Recurrent Unit (GRU), the RawNet uses the raw speech waveforms as an input to extract frame-level speaker embeddings utilizing residual blocks with CNN. In RawNet, the use of a first layer that is completely unconstrained and whose parameters are automatically learned might lead to delayed learning. In particular, when training data is sparse, the first layer outputs frequently exhibit noise. With SincNet, this problem can be addressed because the first convolutional layer uses the raw waveform as input and consists of a bank of bandpass filters with *sinc* function-parametrized parameters. Few parameters, mainly the cut-in and cut-off frequencies with a fixed rectangular-shaped filter response, are learned when a confined first layer is used. It results in the learning of a filterbank structure and outputs that are more meaningful. By substituting the first layer of RawNet with a SincNet layer, RawNet2 combines SincNet and RawNet and takes advantage of the benefits of both methodologies. In this work, we used a RawNet2 architecture that is identical to that described in [240].

5.5.3 Speaker-Microphone Distance-Based Analysis

As done for the Bump, and Morlet wavelet-based analyses, we show the analysis of pop noise strength (measured via its energy) w.r.t. distance between the speaker and the microphone for the case of Morse wavelet-based scalogram. Figure 5.21

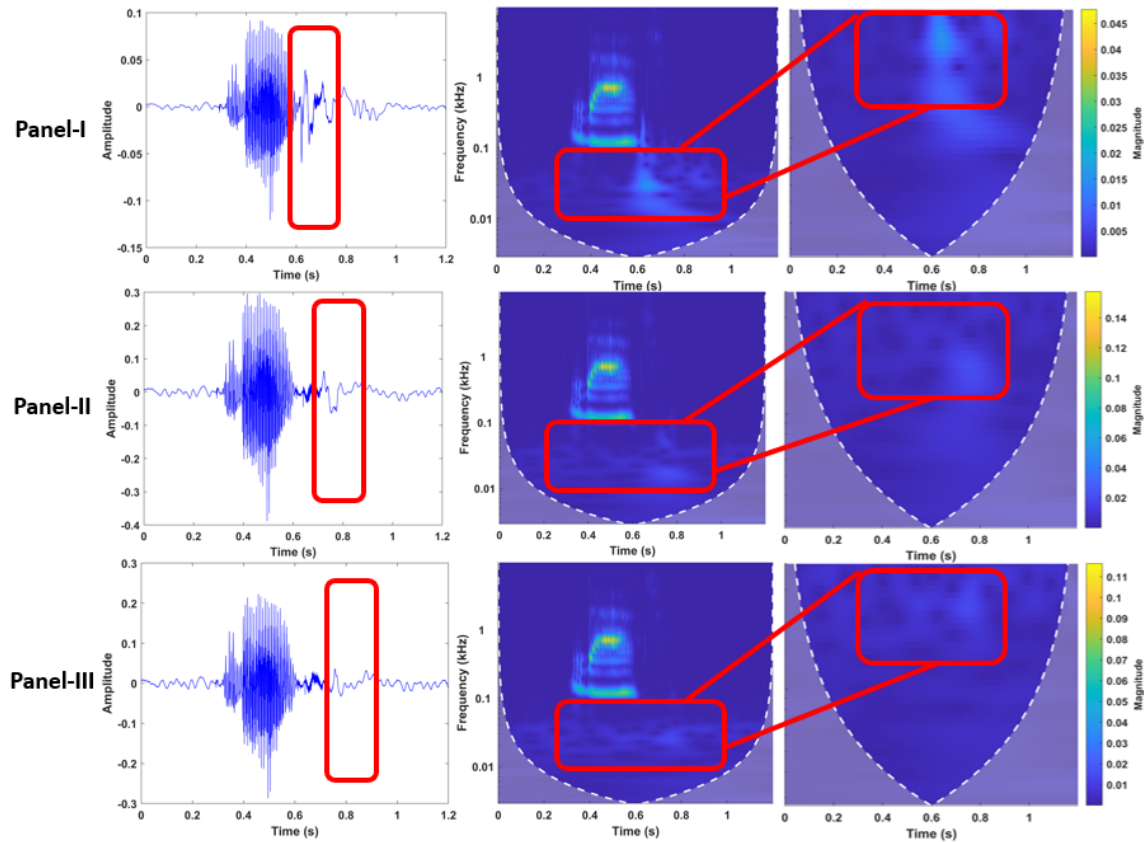


Figure 5.21: Panel I, Panel II, and Panel-III represent the varying distance of the speaker from the microphone, i.e., 5 cm, 5.39 cm, and 6.42 cm, respectively, for (a) time-domain signal for the word ‘dad’, (b) corresponding scalogram, and (c) selected region of scalogram corresponding to low-frequency (0 – 40 Hz). Solid boxes in red indicate the presence of pop noise. Best viewed in color.

shows the capturing of pop noise for three distance values (i.e., 5 cm, 5.39 cm, and 6.42 cm) using Morse wavelet-based scalogram. It can be observed that for the smallest speaker-microphone distance (as shown in Panel-I), the pop noise is distinct, and it becomes less prominent as the distance is increased (as shown in Panel-II and Panel-III).

Given that the strength of pop noise also depends on the type of phoneme uttered, we perform the analysis w.r.t. the different phoneme categories using Algorithm 8, as shown in Figure 5.22, which shows the effect of speaker-microphone distance w.r.t. each of the phoneme types. Using this analysis, we observe the following:

Algorithm 8 Proposed Algorithm for Pop Noise Energy Estimation Using Morse Wavelet for VLD.

```

1: procedure ENERGY_POP( $x$ )                                ▷  $x$  is the speech signal
2:    $w\_name = 'morse'$                                        ▷ Taking Morse wavelet
3:    $[cwt\_coeffs, F] \leftarrow cwt(x, w\_name)$ 
4:    $Low\_bins \leftarrow \text{find}(0 < F \leq 40 \text{ Hz})$ 
5:    $Low\_coeffs \leftarrow cwt\_coeffs(Low\_bins)$ 
6:    $Pop\_energy = [abs(Low\_coeffs)]^2$ 
7:    $[r, c] = \text{size}(Pop\_energy)$ 
8:   for  $i \leftarrow 0 : r$  do
9:      $E\_LF(i) = \text{sum}(Pop\_energy(i, :))$ 
10:  end for                                                ▷ Each row of E_LF has energy for 1 frequency bin
11:   $E_{mean} = \text{mean}(E\_LF)$ 
12: end procedure

```

- Pop noise energy shows a *decaying exponential trend* for plosives, fricatives, whispers, and affricate phonemes, for *all* the energy estimation approaches, i.e., STFT, Morlet wavelet-based, Bump wavelet-based, and the proposed Morse wavelet-based approaches. However, for the cases of nasal and liquids (as shown in Figure 5.22 (e) and (f)), the decaying trend is not observed (rather it is near-flat) with increasing distance, which can be observed from the trendline equations as shown in Table 5.5, where the nasal and liquid phoneme types have a *very small* value of the exponent, i.e., very slow decay. In fact, for the STFT and Bump wavelet methods, liquid phonemes show *no decay* at all, with time constant in the trendline equations as 0.0054, and 0.0321, respectively.
- Given that liquids are semi-vowels [10], they have very less or almost no pop noise. For the case of nasal sounds, the spectrum is dominated by the low resonance of the large volume of the nasal cavity as compared to the oral cavity [10], thereby having dampened impulse response, which in turn means that the nasal cavity has a large -3 dB bandwidth and thus, more energy loss. This can be explained by the design of a 2^{nd} order resonator, the z-plane pole radius and -3 dB bandwidth are related as $r = e^{-\pi BT}$ (w.r.t. impulse-invariant transformation mapping of s-plane pole to z-plane, as discussed in eq. (6.13) Chapter 6), where T is the sampling period, implying that $r \propto 1/B$. Further, a quick damping of the impulse response in the nasal cavity implies $|r| \ll 1$, and thus, B being large, the energy losses are more. Therefore, in the case of nasal sounds, the impulse response itself is dominantly present in the lower frequency region. Given that pop noise

and the system response are *both* in the lower frequency regions, it is thus difficult to detect the presence of pop noise in nasal sounds.

- Keeping the distance fixed, say at 5 cm, it can be observed that plosives and fricatives have similar amount of pop noise energy, followed by whisper and affricate sounds. Furthermore, nasal and liquids have little pop noise energy as compared to the rest of the phoneme sounds.
- Morse wavelet captures lower energy as compared to the Morlet wavelet for all the distances, indicating that for the case of replay attack (i.e., when distance is large enough), Morse wavelet is more effective than Morlet and hence, it also gives better performance in terms of classification accuracy, as discussed in subsection 5.5.4.4.
- It is also observed that STFT-based method is not able to capture pop noise effectively even for smaller distances. This is because the basis function used for STFT is $\{w(t)e^{j\omega t}\}_{t \in R}$, where it comprises cosine and sine functions and window $w(t)$ of larger duration of 20-30 ms. However, pop noise is transient in nature. Therefore, STFT is not able to capture pop noise effectively, as compared to wavelet-based approaches, which utilize wavelets of *limited duration, i.e., relatively smaller support* as basis functions.

Table 5.5: Trendline Equations (in the form of amplitude and time constant (a, b)) Obtained for Each Method w.r.t. Phoneme Type

Phoneme Type	Method			
	STFT	Bump	Morlet	Morse
Plosives	0.2542, -0.204	0.4458, -0.233	0.7738, -0.228	0.643, -0.236
Fricatives	0.344, -0.281	0.6011, -0.282	0.9689, -0.262	0.6007, -0.221
Whisper	0.264, -0.251	0.4484, -0.254	0.8154, -0.237	0.7036, -0.262
Affricates	0.2027, -0.211	0.3565, -0.227	0.6161, -0.223	0.4947, -0.219
Nasal	0.0357, -0.015	0.0506, 0.0046	0.0977, -0.00008	0.7036, -0.262
Liquids	0.028, 0.0054	0.0364, 0.0321x	0.1452, -0.033	0.0659, -0.014

5.5.4 Experimental Results

5.5.4.1 Effect of $P_{\beta,\gamma}^2 = \beta\gamma$

The shape and size of the Morse wavelet are controlled by two parameters, namely, $P_{\beta,\gamma}^2$ and γ . The parameter $P_{\beta,\gamma}^2$ (as shown in eq. (5.10)) is also called the wavelet duration, which controls the bandwidth of the mother wavelet function [9]. To that effect, we perform experiments on the Dev and Eval sets of the POCO dataset, using CNN as the classifier. In order to observe the effect of $P_{\beta,\gamma}^2$, we keep the

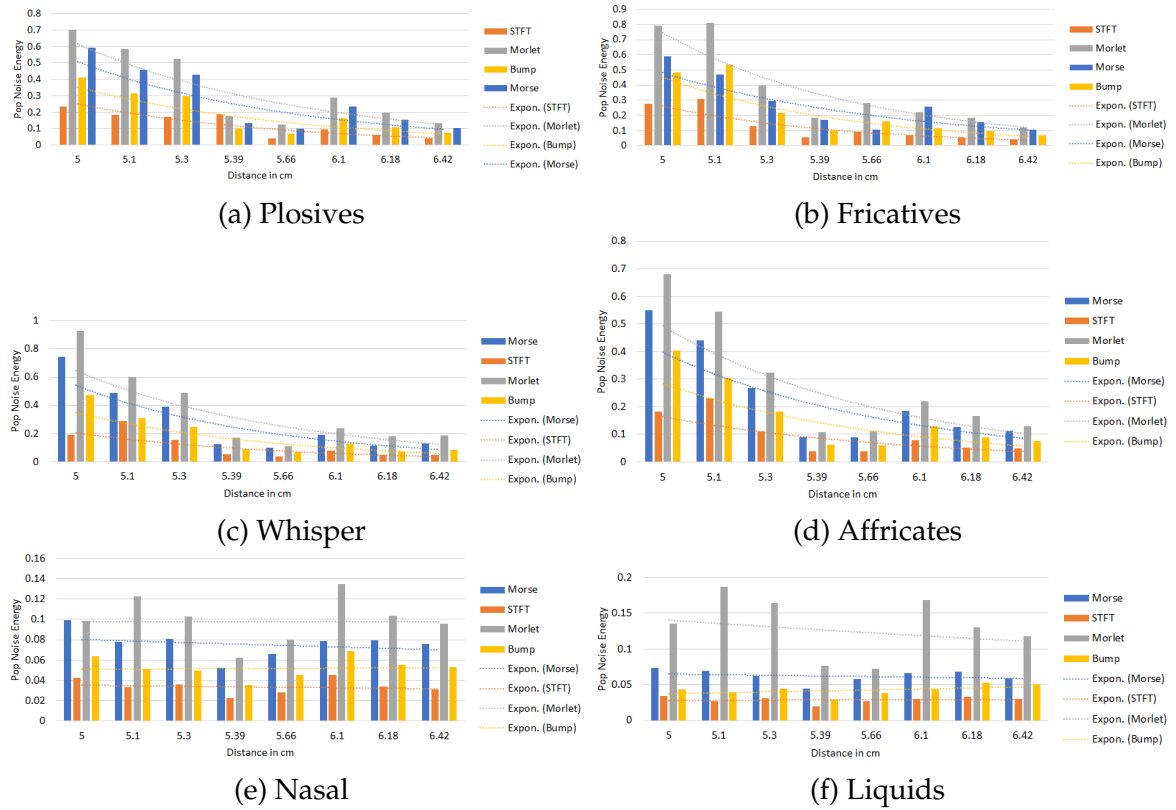


Figure 5.22: Pop noise energies of various phonemes plotted w.r.t. the distance of the speaker from various microphones for the case, when the speaker is at a distance of 5 cm from Mic 7. The pop noise energies are obtained using the proposed Algorithm 8. The trendlines in each of the sub-figures indicate that the energy of pop noise decreases with the distance of the speaker’s mouth from the microphone.

value of wavelet parameter γ to be fixed as 3 (since the analysis in Figure 5.18 shows that the optimal concentration of energy is obtained for $\gamma = 3$). Figure 5.23 shows the performance in terms of percentage accuracy, when $P_{\beta,\gamma}^2$ is varied. It can be observed that the optimal value of $P_{\beta,\gamma}^2$ obtained is 6, which gives us the accuracy of 90.15% and 87.01% on the Dev and Eval sets, respectively. For the rest of the experiments in this chapter, the parameter $P_{\beta,\gamma}^2$ is kept fixed as 6.

5.5.4.2 Effect of γ

After obtaining the optimal value of $P_{\beta,\gamma}^2$ as 6, experiments are performed by varying the parameter γ . The value of γ is such that it satisfies the constraints of $P_{\beta,\gamma}^2/\gamma \leq 40$, and $P_{\beta,\gamma}^2 > \gamma$ as given in the freely available MATLAB toolbox called JLAB, available at <http://www.jmlilly.net> [8]. To that effect, the performance of the VLD system with CNN-based classifier is shown in Figure 5.24. We find that the optimal value of γ is 3, which gives accuracy of 90.15% and 87.01%,

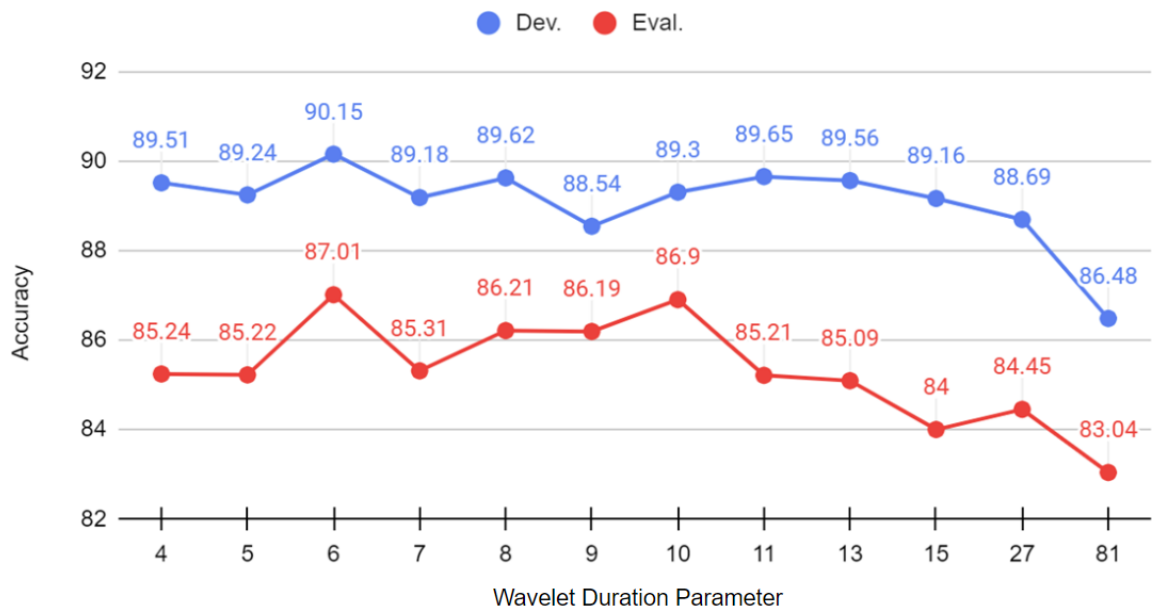


Figure 5.23: Results (in % Accuracy) for the proposed Morse wavelet-based feature set on the Dev and Eval set of POCO dataset, to observe the effect of wavelet duration parameter (i.e., $P_{\beta,\gamma}^2$).

on the Dev and Eval sets, respectively. The obtained optimal value of γ as 3 is in

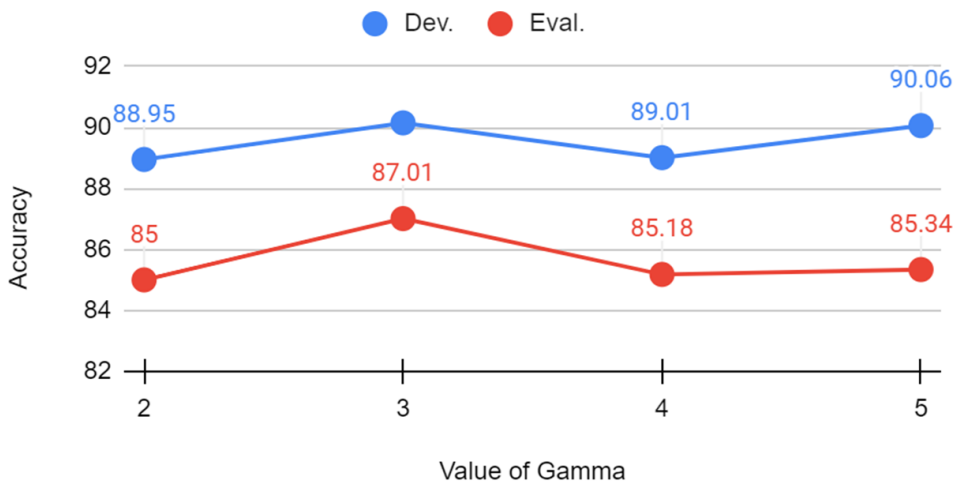


Figure 5.24: Results (in % Accuracy) for the proposed Morse wavelet-based feature set on the Dev and Eval set of POCO dataset, in order to observe the effect of γ parameter.

agreement with the fact that the wavelets for $\gamma = 3$ for a fixed $P_{\beta,\gamma}^2$ are known to be optimal as the value of $\tilde{\Psi}_{3;\beta,\gamma}(\omega_\psi)$ from eq. (5.11) is 0. The quantity $\tilde{\Psi}_{3;\beta,\gamma}(\omega_\psi)$ controls the skewness of the Morse wavelet and for $\gamma = 3$ the skewness vanishes, making the Morse wavelet to be strictly symmetrical in shape.

5.5.4.3 Effect of Frequency Range

To observe the effect of frequency range on the performance, we performed experiments in the lower frequency regions, by increasing the bandwidth of the frequency range by a step size of 10 till 80 Hz. To that effect, CNN is used as the classifier, and RC-A and RP-A subsets of the POCO dataset are used, with their details in Section 3.3.1 in Chapter 3. The Morse wavelet parameters were fine-tuned at $P_{\beta,\gamma}^2 = 6$, and $\gamma = 3$. Our experiments showed that the best performance is achieved for the frequency range of 1 – 50 Hz. Furthermore, to validate the presence of pop noise in low frequencies (i.e., 1 – 50 Hz), we performed further set of experiments on the *complementary* frequency range, to show the absence (or weak presence) of pop noise. Notably, from Table 5.6, it can be observed that the highest VLD accuracy is obtained on 1 – 50 Hz frequency range, and likewise the lowest VLD accuracy is obtained on 50 – 11025 Hz frequency range, indicating a strong evidence of presence of pop noise in the lower frequency range of 1-50 Hz.

Table 5.6: Results (in % Classification Accuracy) for Morse-CNN-Based Pop Noise Detection Method with Variation in Frequency Range

Freq. Range	Dev	Eval	Freq. Range	Dev	Eval
1-10 Hz	88.46	84.65	10-11025 Hz	79.60	74.86
1-20 Hz	88.72	84.84	20-11025 Hz	77.41	70.19
1-30 Hz	89.56	84.72	30-11025 Hz	73.89	66.74
1-40 Hz	90.15	87.01	40-11025 Hz	72.40	64.69
1-50 Hz	90.55	88.43	50-11025 Hz	58.56	54.65
1-60 Hz	89.30	86.46	60-11025 Hz	70.60	61.81
1-70 Hz	89.45	86.21	70-11025 Hz	70.94	62.24
1-80 Hz	90.53	87.60	80-11025 Hz	68.03	60.69

5.5.4.4 Phoneme-Based Analysis

As described in Section 3.3.1 of Chapter 3, the utterances in the POCO dataset are divided into various phoneme types. Figure 5.25 shows word-wise VLD accuracy of all the 44 words in the dataset. Correspondingly, one example from each phoneme type is taken (i.e., ‘tip’ for plosive, ‘who’ for whisper, ‘laugh’ for fricative, ‘chip’ for affricate, ‘arm’ for nasal, and ‘run’ for liquid phoneme types) and the corresponding scalograms are analysed in Figure 5.26 and Figure 5.27, for genuine and spoofed replay cases, respectively. Furthermore, Table 5.7 shows the experimental results obtained for various phoneme classes using the RC-A and RP-A subsets of the POCO dataset, using CNN as the classifier. It can be observed

that the proposed Morse wavelet-based method achieves relatively the best performance for most of the phoneme types, such as plosives, fricatives, affricates, and liquids. High performance accuracy of 89.72% and 90.61% is achieved for plosives, and fricatives, respectively. This indicates that pop noise is more likely to be present in words having plosives and fricative phonemes. Further, it can be observed that pop noise is likely to present more dominantly in plosives as compared to the fricative sounds. We believe that this may be due to the fact that the production of the plosive sounds involves four events (Section 3.4 in Chapter 3 in [10]) - (1) first is the complete closure of the oral tract when no sound is radiated from the lips and there is air pressure buildup in the oral cavity, (2) this is followed by formation of turbulence due to the release of air pressure over a very short duration i.e., *burst* (or impulsive) source, (3) the air rushes through the open oral cavity leading to generation of aspiration due to the turbulence at the open vocal folds (i.e., before the onset of the vocal folds vibration), as air rushes through the open oral cavity after the burst (4) there is onset of the vowel sound about 40 – 50 ms after the burst for unvoiced plosives. Perceptual experiments indicate that if the release of the burst and the onset of voicing are between 20 ms of each other, the consonant is considered voiced, otherwise it is categorized as unvoiced. Thus, the buildup leading to the "burst" indicates the chances of breath sounds (i.e., pop noise) to be captured by the microphone.

Table 5.7: Phoneme-wise Average VLD Accuracy (in %)

Phoneme Type	STFT-based baseline [103]	Bump wavelet-based method [92]	Morlet wavelet-based method [28]	Proposed Morse wavelet-based method
Plosives	71.72	81.58	89.07	89.72
Whisper	76.83	81.09	86.21	85.50
Fricatives	75.55	80.77	87.61	90.61
Affricates	71.83	78.53	85.26	85.83
Nasal	59.33	76.50	80.77	76.50
Liquids	56	69.87	79.49	86

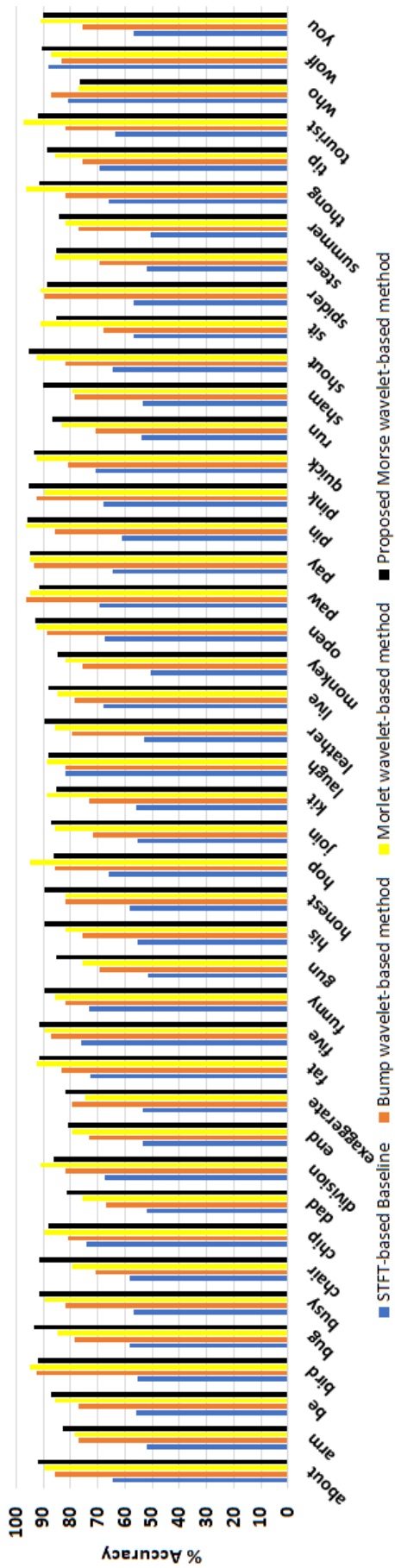


Figure 5.25: Word-wise VLD accuracies (in %) for the various existing methods compared with the proposed Morse wavelet-based method.

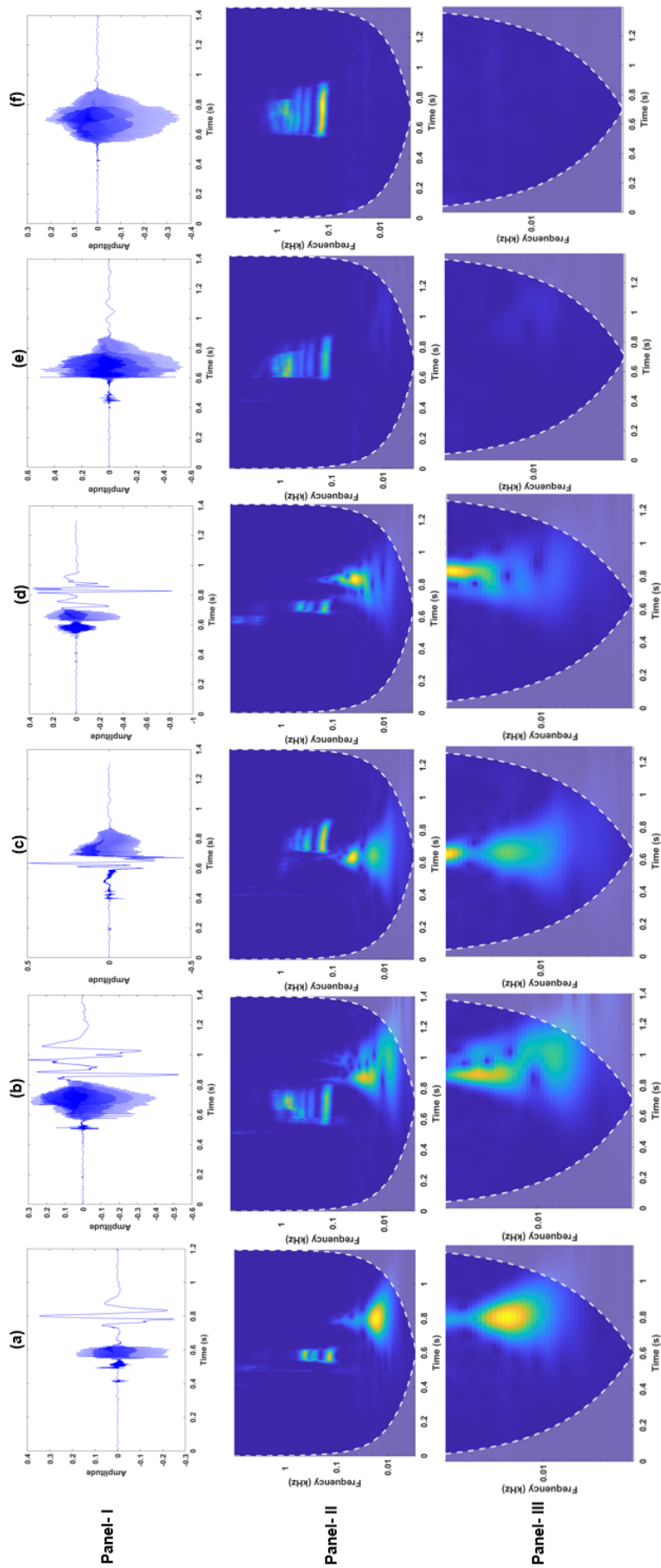


Figure 5.26: Phoneme-wise analysis of genuine (live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morse wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is present, for (a) plosive (the sample word is 'tip'), (b) fricative (the sample word is 'laugh'), (c) whisper (the sample word is 'who'), (d) affricate (the sample word is 'chip'), (e) nasal (the sample word is 'arm'), and (f) liquid (the sample word is 'run').

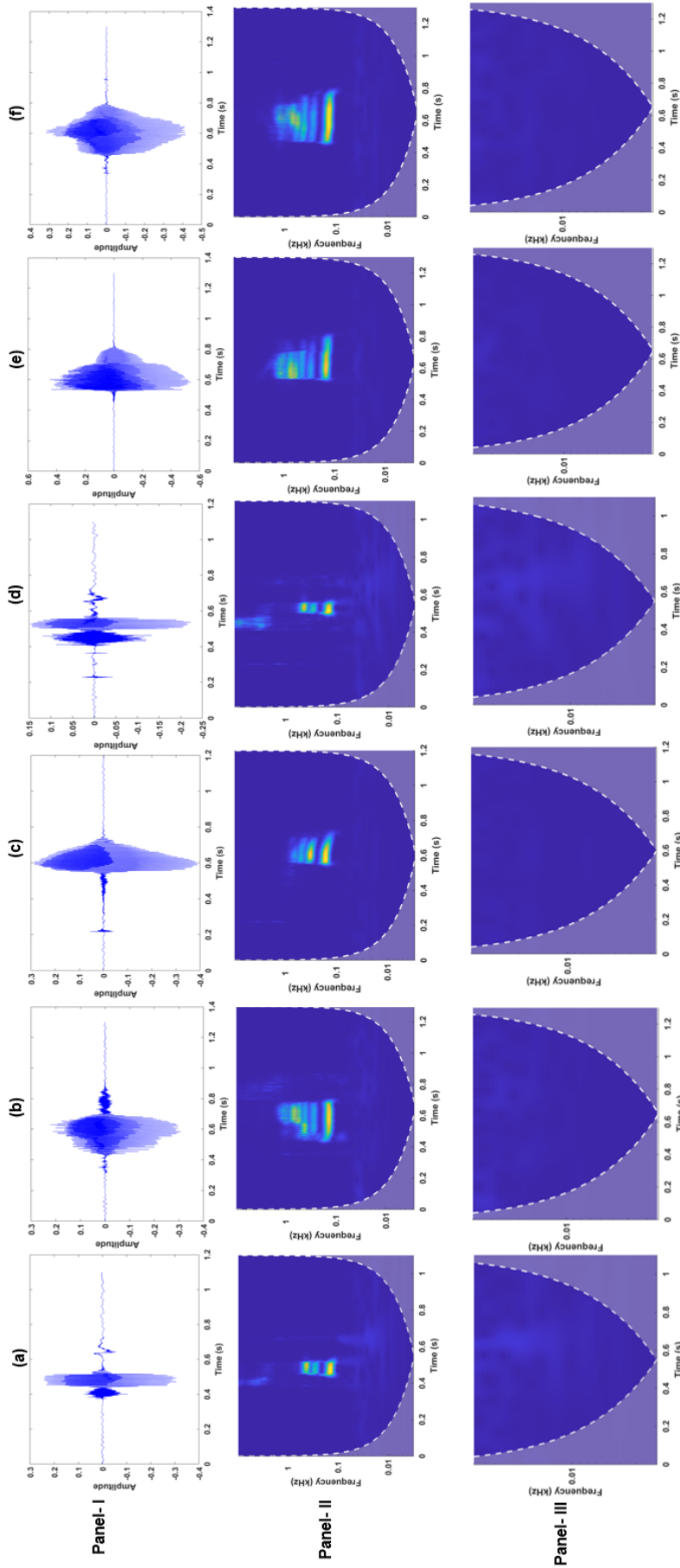


Figure 5.27: Phoneme-wise analysis of spoofed (non-live) speech, where Panel-I shows the time-domain signal, Panel-II shows the Morse wavelet-based scalogram, Panel-III shows the selected region of scalogram where pop noise is absent, for (a) plosive (the sample word is 'tip'), (b) fricative (the sample word is 'laugh'), (c) whisper (the sample word is 'who'), (d) affricate (the sample word is 'chip'), (e) nasal (the sample word is 'arm'), and (f) liquid (the sample word is 'run').

5.5.4.5 Effect of Distance Between Genuine Speaker and Attacker’s Microphone

REP-A subset contains simulated replay utterances with varying distances between the genuine speaker, and the attacker’s microphone. In this work, three cases of the distances between the attacker and the speaker are considered as 30 cm, 50 cm, and 70 cm, on which the performance of VLD system is evaluated on the REP-A subset. To that effect, subband-wise evaluation is done in order to observe the relative significance of a particular frequency subband in the performance. To that effect, Table 5.8 shows the subband-wise VLD accuracies for the 3 distances of REP-A subset. It can be observed that % classification accuracy shows

Table 5.8: Speaker and Attacker Distance-wise Performance (in % Accuracy) for Morse Wavelet-Based VLD using RC-A (genuine) *vs.* REP-A (spoo) Dataset with Variations in Subband Frequency Range with CNN as the Classifier.

	30 cm		50 cm		70 cm	
Subband	Dev	Eval	Dev	Eval	Dev	Eval
1 to 10 Hz	96.91	94.72	96.38	94.37	96.79	94.90
1 to 20 Hz	95.65	94.00	96.44	94.50	96.73	94.90
1 to 30 Hz	97.66	96.31	97.20	95.36	97.02	95.31
1 to 40 Hz	97.66	96.31	97.29	95.84	95.89	94.80
1 to 50 Hz	98.13	98.00	97.93	97.09	98.36	96.81
1 to 60 Hz	97.93	96.68	96.15	95.21	96.79	96.04
1 to 70 Hz	97.75	97.25	96.09	95.04	96.91	95.83
1 to 80 Hz	97.52	96.93	96.62	96.01	96.72	95.25

an increasing trend from 1 to 50 Hz. In particular, for 1 to 50 Hz subband, the performance is the highest for *all* the three cases of distances between the genuine speaker and the attacker’s microphone (i.e., 30 cm, 50 cm, and 70 cm). Furthermore, as we go for higher subbands, we observe a decrease in the performance of the VLD system. Therefore, Table 5.8 validates that pop noise is present in low frequency regions, particularly between 1 to 50 Hz. To that effect, the rest of the experiments in this work w.r.t. Morse wavelet-based features are performed by considering 50 Hz as the optimal benchmark for pop noise detection.

It can also be observed from Table 5.8 that for most of the subbands, the performance accuracy decreases with increase in speaker-attacker distance. In particular, for the case of frequency range of 1 to 50 Hz, the Eval accuracies are 98%, 97.09%, and 96.81% for the distances of 30 cm, 50 cm, and 70 cm, respectively.

5.5.4.6 Effect of Classifier Structure

In this subsection, we observe the effect of different classifiers for the VLD task. To that effect, experiments are performed using CNN, LCNN, and ResNet classifiers, the details of which are described in detail in subsection 5.5.2. Table 5.9 shows

Table 5.9: Overall Performance (in terms of % Accuracy) of Various Feature Sets Across Three Different Classifiers, Namely, CNN, LCNN, and ResNet.

Feature Set	Classifier	Accuracy	
		Dev	Eval
STFT	CNN	70.57	71.81
	LCNN	70.60	71.90
	ResNet	72.05	71.84
CQT [93]	CNN	81.52	81.82
	LCNN	84.84	82.45
	ResNet	83.04	80.42
Bump Wavelet	CNN	78.08	75.03
	LCNN	75.99	73.59
	ResNet	74.56	71.43
Morlet Wavelet	CNN	87.26	86.23
	LCNN	85.31	82.30
	ResNet	87.61	83.53
Morse Wavelet	CNN	90.55	88.43
	LCNN	88.86	86.74
	ResNet	91.02	88.33

the performance of each of the feature sets, compared with the proposed Morse wavelet-based features, for all the three classifier structures. For Morse wavelet-based features, it can be observed that among all the classifiers, the CNN gives the highest VLD accuracy of 88.43% on Eval set. Furthermore, Morse wavelet outperforms all the remaining feature sets. This is due to the generalizability and *strictly* analytic behaviour of GMWs (as discussed in subsection 5.5.1.2).

Further experiments are performed by considering the REP-A subset, and the different classifiers using the Morse wavelet-based features. To that effect, the three distances of 30 cm, 50 cm, and 70 cm between the attacker and the speaker are considered, and performance accuracy is shown w.r.t. the three classifiers for the Morse wavelet features with frequency range of 0 to 50 Hz. Table 5.10 shows the performance of each of the classifiers w.r.t. speaker-attacker distance. It can be observed from the accuracies on the Eval set, that as the distance increases, the performance of the VLD system decreases, for all three cases of the classifiers.

Table 5.10: Speaker and Attacker Distance-wise Performance (in % Accuracy) for Morse Wavelet-Based VLD Using RC-A (genuine) *vs.* REP-A (spoof) Dataset with Variations in Classifier Structure.

	30 cm		50 cm		70 cm	
Classifier	Dev	Eval	Dev	Eval	Dev	Eval
CNN	98.13	98.00	97.93	97.09	98.36	96.81
LCNN	96.70	96.53	97.31	96.48	97.11	96.12
ResNet	98.28	98.34	98.45	97.98	98.28	97.36

5.5.4.7 Performance Under Ideal Conditions

To evaluate the performance under ideal scenarios, we performed experiments on the POCO dataset for 2 scenarios: (1) when the system is not under attack, and (2) when it is under attack. The scenario where the system is not under attack, means that the inputs to the SSD system are not spoofed signals, i.e., they are strictly genuine signals. Therefore, the performance is evaluated by taking genuine utterances. To do so, the VLD system with Morse wavelet-based features and CNN as the classifier was tested on only genuine utterances. An ideal system will accept all the genuine utterances. Our experiment shows that the proposed VLD system yields 88.30% accuracy, as shown in Table 5.11 when only genuine utterances are given to the system.

For case 2, when the system is under attack, we performed an experiment where only spoofed utterances were given to the system. An ideal system will reject all the utterances. For the proposed system, it was shown that the system rejected 89% of the spoofed utterances, indicating further scope for improvement.

Condition	% Accuracy
Not under attack	88.30
Under attack	89

Table 5.11: Performance Under Ideal Conditions.

5.5.4.8 Performance Comparison With End-To-End Neural Network Model

Further experiments were also performed by considering RawNet2, which is an end-to-end classifier [240]. Table 5.12 shows that the proposed system performs much better than end-to-end model, indicating the significance of handcrafting features using signal processing approaches over direct end-to-end deep learning models, where raw audio is fed as input. More so, the proposed Morse wavelet-

Table 5.12: Overall performance (in terms of % Accuracy on the Eval set) of the proposed system compared with end-to-end RawNet2 model.

VLD system	% Accuracy
End-To-End RawNet2	73.39
Proposed Morse wavelet features with CNN as the classifier	88.43

based features outperform the RawNet2 model by a significant absolute difference of 15.04%.

5.6 Chapter Summary

In this chapter, three analytic wavelet-based feature sets using Bump, Morlet, and Morse wavelets are developed for the VLD task. The chapter discussed the significance of CWT for the VLD task. More so, for each of the three wavelets discussed in this chapter, experimental results and distance-based analysis are shown w.r.t. pop noise detection. In particular, GMWs are shown to be a superfamily of analytic wavelets, and hence, much detailed experiments and analysis are shown w.r.t. Morse wavelet-based features for the VLD task. However, one of the limitations of this work is that the distance between the speaker and the microphone is considered to be fixed without estimating the variations in distance caused due to movement of the speaker’s articulators, as well as the head movements. To that effect, for precise estimation of distance values, source localization techniques can be explored in future and hence, remains an open research issue.

Given that VLD relies on the characteristics of live speech (i.e., pop noise), instead of relying on the characteristics of spoofed speech, it is a step towards anti-spoofing, which is *independent* of the type of spoofing attack. Therefore, such approach has a scope for generalizability of CM systems. Given the attacker has the freedom to mount any type of attack, generalizability of CM system is crucial for practical deployment. To that effect, the next chapter discusses some of the attacker’s perspectives w.r.t. voice privacy, which aims at hiding a speaker’s identity while keeping linguist content and naturalness intact.

CHAPTER 6

Voice Privacy and Attacker's Perspective

6.1 Introduction

The need for privacy-preservation in ASV systems is another important concern. Apart from spoofing attacks, speaker data forgery is also prevalent, which further puts the ASV system's security at serious risk [114]. In a speech signal, apart from the linguistic content, there are traits of the speaker, such as accent, pitch (or fundamental frequency F_0), tone, rhythm, and idiosyncrasies. Hence, considering an individual's identity, it is not just his/her name, but it is also the other traits that are captured by the speech signal, such as gender, age, health status, personality, emotional state, and accent. So far as the practical deployment of ASV technology is concerned, designing a privacy preserving system for speakers' identities (i.e., Voice Privacy (VP) system) is crucial. A VP system is designed to preserve the privacy of users, without altering the linguistic content. A VP system can be used for real-world applications, such as in forensics, in voice biometric systems, in medical-domain and to study and analyze attacker's perspectives to build more secure ASV systems as shown in Figure 6.1.

This chapter ¹ discusses VP and attacker's perspective. Recently, efforts were

¹This Chapter is based on the following publications:

- **Priyanka Gupta**, Hemant A. Patil, and Rodrigo Capobianco Guido "On Vulnerability Issues in Automatic Speaker Verification (ASV) Systems", accepted (under minor revision) in EURASIP Journal on Audio, Speech, and Music (JASM) Processing, Special Issue on Security & Privacy in Speech Communication, 2023, 21 pages.
- **Priyanka Gupta**, Gauri P. Prajapati, Shrishti Singh, Madhu R. Kamble and Hemant A. Patil, "Design of Voice Privacy System using Linear Prediction," in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, December 7-10, 2020, pp. 543-549.
- **Priyanka Gupta**, and Hemant A. Patil, "Voice Biometrics: Attacker's Perspective," Gerard Chollet, and Carmen Garcia Mateo (Eds.) in Voice Biometrics: Technology, trust and security, Institution of Engineering and Technology (IET), 2021.
- **Priyanka Gupta**, Shrishti Singh, Gauri P. Prajapati and Hemant A. Patil, "Voice Privacy in Biometrics," in Biomedical Signal and Image Processing with Artificial Intelligence, 2023.

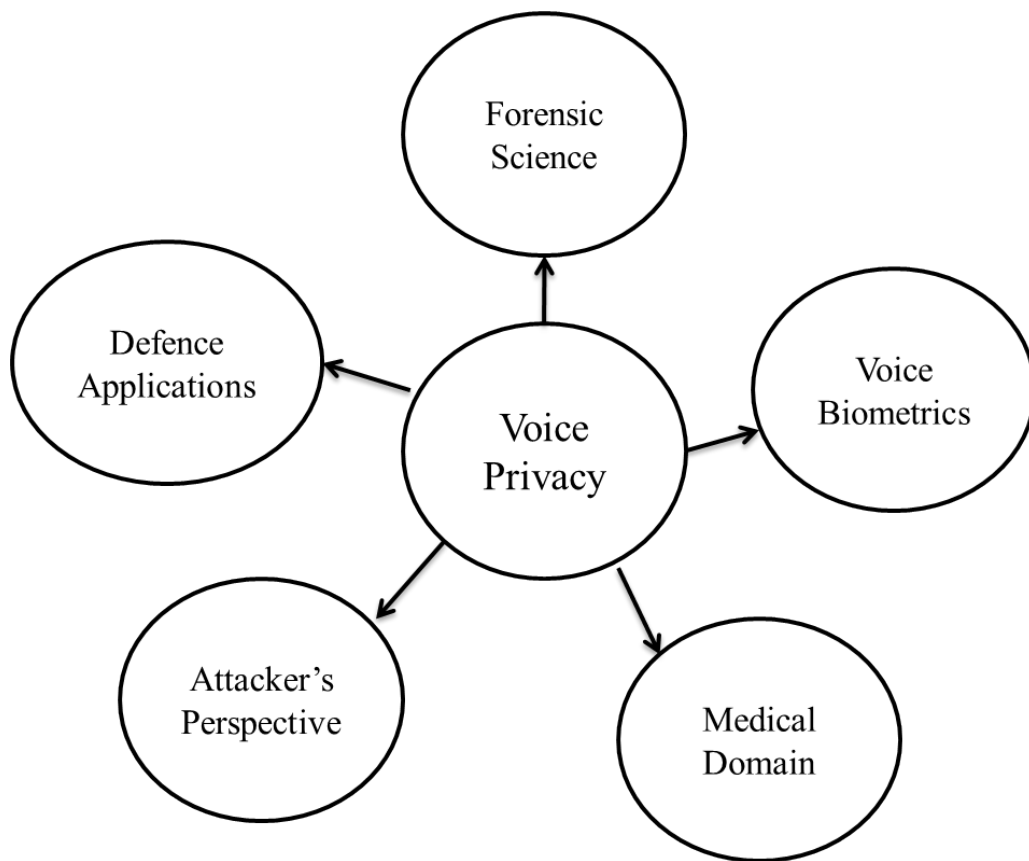


Figure 6.1: Applications of Voice Privacy.

made to develop privacy preservation solutions for speech technology. In the light of moving forward towards this development, the first VP challenge was being organized during INTERSPEECH 2020, followed by Voice Privacy workshop during Speaker Odyssey 2020 to motivate researchers in this direction [13, 14].

6.1.1 Motivation for Voice Privacy

The notion of privacy in the field of healthcare is very old. With the advancement in technology, comes the easy data collection and processing technologies [241]. At the same time, the detail and diversity of information collected in the context of biomedical research is increasing at an unprecedented rate. The easy availability of such large amounts of data has also raised the concerns of privacy invasions [242]. It is important to understand the scope and frequency of these invasions. There are cases where medical records of people are illegally accessed for the purpose of identity fraud. Due to privacy concerns, people change the behavioural activities, such as visiting another doctor for check-up, not seeking

care when needed in order to avoid disclosure of medical information, self treating or medicating themselves, not disclosing full information about their medical history, paying out of pocket despite being insured, hesitate to participate in the surveys, which require data from people in the fear of data getting misused, etc. This privacy protective behaviour shows the trust issues of people. Therefore, focus on the privacy preservation technologies should be given utmost importance to reduce the vulnerability of the data. It becomes all the more important in the case of patients suffering from speech disorders, and diseases like dysarthria, which affects the characteristics of the natural speech production mechanism. In such cases, the medical practitioner may have to record and save the patient's speech data (with patients' consent). However, the risk of availability of patients' unprotected speech data will exist. Moreover, this risk will turn severely damaging if the patient is enrolled as a genuine speaker on a voice biometric (or ASV) system. This risk can be mitigated to a large extent if voice privacy measures are applied to the speech data.

6.1.2 De-identification vs. Anonymization

Unwanted users may employ VC and SS techniques to impersonate speakers, when voice is transmitted over the Internet. As a result, it becomes necessary to conceal the speaker's identity from speech recordings. De-identification is the process by which a data custodian modifies or removes an individual's identifying information from a dataset, making it impossible for attackers to identify the subject from whom the speech data was gathered while allowing sharing and reuse of the speech data.

A VP system generates a speech utterance, where the original speaker's identity is hidden or removed. VP system synthesizes a speech signal which has the speaker's identity hidden, without affecting the linguistic content. Therefore, a VP system ensures privacy by transforming a speaker's voice to a *pseudo-speaker's* voice, and hence protecting the privacy of the original speaker. However, this process can be reversible or irreversible depending on the technique used for VP and also depending on the application. For example, reversibility may not be required in some applications, such as those used for Automatic Speech Recognition (ASR) and public environment monitoring, or it may potentially pose a threat to user privacy. The practice of making data anonymous, when reversibility requirements are not met is known as *anonymization* [13]. This means that the identity transformation in anonymization is an irreversible function and hence, it is impossible to reclaim the original identity. On the other hand, de-identification procedures

are reversible in nature and hence, the original identity can be recovered from the pseudo-identity.; given a de-identified spoken utterance, the original speech can be recovered if the necessary parameters are supplied. This usually requires the knowledge of some extra (additional) information, such as a *key*. Therefore, de-identification methods are *generally* based on cryptographic methods.

VP can be achieved by Voice Transformation (VT) techniques, which retain the quality of speech to a certain extent [33]. The VT approaches usually include anonymization by VC, SS, and other techniques of speech processing. One such speech processing technique is using linear prediction (LP) of speech, which is discussed in detail in this chapter. The terms anonymization and de-identification are used interchangeably in this chapter unless stated otherwise.

6.2 Voice Privacy Using Linear Prediction (LP) Model

6.2.1 Speech Production Model w.r.t. Linear Acoustics

Depending on the shape and structure in the time-domain, speech signal can be classified into *voiced* and *unvoiced* speech. Voiced sounds are produced due to quasi-periodic vibrations of the vocal folds. These vibrations occur because of the pressure from the lungs. The contraction of the lungs first results in air flowing through the glottis. According to the Bernoulli's Principle, as the airflow velocity increases, local pressure in the region of the glottis decreases, and the tension in the vocal folds increases. The decrease in pressure and increase in vocal fold tension cause the vocal folds to close shut abruptly. The air pressure then builds behind the vocal folds as the lungs continue to contract, forcing the folds to open. The entire process then repeats, and the result is periodic "puffs" of air that enter the vocal tract. One can actually touch and feel the vibrations of the vocal folds by placing a thumb near the throat while uttering a voiced sound like a vowel (eg., /a/). However, in the case of unvoiced speech, such as /h/, one does not feel any vibrations of the vocal folds (also called *aspiration*, i.e., turbulence created at the vocal folds). This is because, for unvoiced sounds, the vocal folds are just slightly open and therefore, the air rushing from the lungs produces turbulence at the vocal folds. This turbulence is modelled as a noisy signal as shown in Figure 6.2, which shows discrete-time speech production model for voiced and unvoiced sounds. For voiced sound, the gain is A_v , which corresponds to the loudness. Similarly, A_N corresponds to the loudness of the unvoiced sound. Considering the voiced case, the overall transfer function of the speech production model is

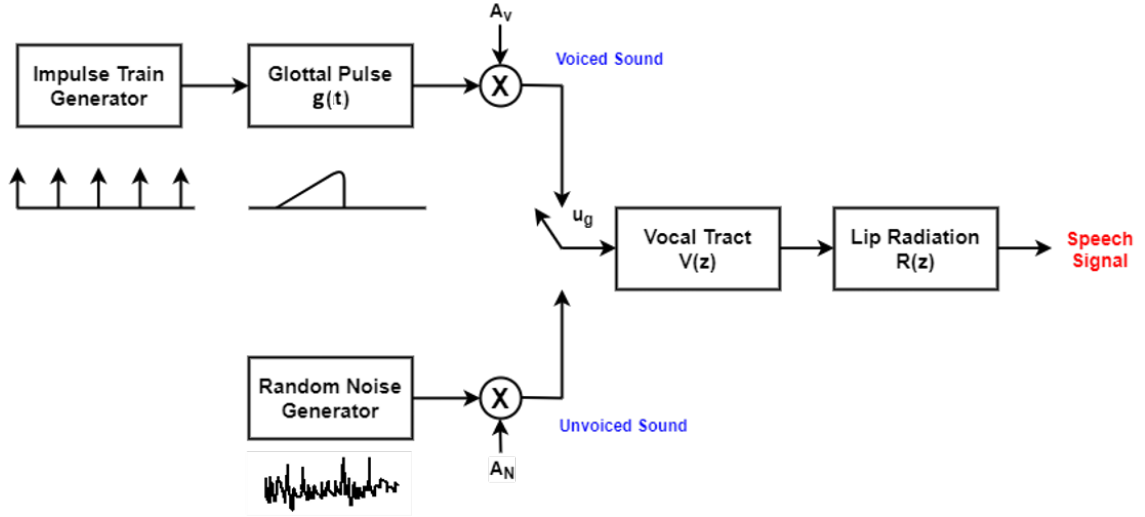


Figure 6.2: Discrete-Time Speech Production Model. After [10].

$H(z) = G(z)V(z)R(z)$, where $G(z)$ is the transfer function of the glottal system, $V(z)$ is the transfer function of the vocal tract system, and $R(z)$ is the lip radiation effect. $G(z)$, $V(z)$, and $R(z)$ represents z -domain system functions. Mathematically, $G(z)$ is given by [10];

$$G(z) = \frac{1}{(1 - c_k z^{-1})^2}, \quad (6.1)$$

where $(1 - c_k z^{-1})^2$ and $(1 - c_k^* z^{-1})^2$ are complex conjugate poles with $|c_k| < 1$ [10]. Furthermore, the vocal tract system $V(z)$ and lip radiation $R(z)$ are given by [10]:

$$V(z) = \frac{G}{\prod_{k=1}^{N/2} (1 - 2r_k \cos \theta_k z^{-1} + r_k^2 z^{-2})}, \quad (6.2)$$

$$R(z) = R_o(1 - z^{-1}), \quad (6.3)$$

where G is the gain of $V(z)$, r_k and θ_k are the pole radius and pole angle, respectively, of the k^{th} complex pole-pair. If $e^{-cT} \approx 1$, then $H(z)$ will be

$$H(z) = \frac{\sigma}{1 - \sum_{k=1}^p a_k z^{-k}}, \quad (6.4)$$

where σ is the gain of $H(z)$. The vocal tract system $V(z)$ is modelled as a linear time-invariant (LTI) all-pole system by cascading the 2^{nd} order digital resonators corresponding to each complex pole-pair and thus, formants as shown in Figure 6.3. As per L. G. Kersta, who reported one of the first studies in speaker recognition, resonance is defined as *reinforcement* of spectral energy at or around a particular frequency [243]. The resonance frequencies of the vocal tract system are called formant frequencies, which implicitly capture the shape of the vocal tract

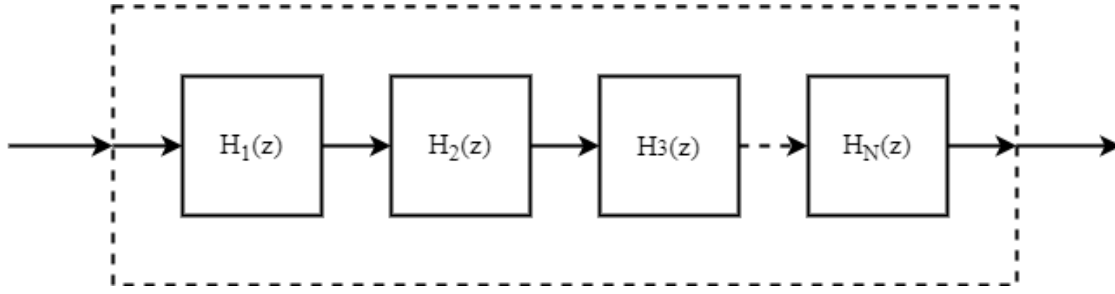


Figure 6.3: Vocal tract system, $V(z)$ modelled by cascading 2^{nd} order digital resonators. After [10,11].

system and, thus, form the spectrum. The peaks in the spectrum are referred to as formant peaks. The acoustics of the vocal tract are usually sculptured employing a mathematical all-pole model. The frequencies of the poles of this filter model fall near those of the formant. A formant is also referred to as the acoustic resonance of the human vocal tract. The spectral peaks of the spectrum are referred to as formant. The formants change with different shapes, and sizes of vocal tract configurations [244]. The formants F1-F4 are the first four lowest resonant frequencies of the vocal tract. Vowels typically have four or more distinguishable formants, and sometimes more than six, though most vowels can be characterized by F1 and F2. The formant F3 helps to differentiate between [i] and [y] -like sounds. Additionally, plosives (and, to some degree, fricatives) modify the placement of formants in the surrounding vowels, and that is where F3 and F4 come into play. Therefore, the vocal tract system by cascading the four 2^{nd} order digital resonators (corresponding to the first four formants) is given by

$$V(z) = \prod_{i=1}^4 H_i(z), \quad (6.5)$$

where each $H_i(z)$ is a 2^{nd} order digital resonator, as shown in Figure 6.3. The transfer function for 2^{nd} order digital resonator for i^{th} formant is given by:

$$H_i(z) = \frac{1}{(1 - p_{1_i}z^{-1})(1 - p_{2_i}z^{-1})}, \quad (6.6)$$

where p_1 and p_2 are the complex conjugate pole-pair of 2^{nd} order resonator transfer function. For i^{th} formant, $p_{1_i} = p_{2_i}^* = r_i e^{\pm j\omega_{o_i}}$. Taking Discrete-Time Fourier Transform (DTFT) of $H_i(z)$ frequency response of i^{th} formant is given by:

$$H_i(z)|_{z=e^{j\omega}} = H_i(e^{j\omega}) = \frac{1}{(1 - r_i e^{j\omega_{o_i}} e^{-j\omega})(1 - r_i e^{-j\omega_{o_i}} e^{-j\omega})}, \quad (6.7)$$

where ω_{o_i} is the pole angle and r_i is the pole radius for the i^{th} formant. Now, taking magnitude of $H_i(e^{j\omega})$, we get,

$$|H_i(e^{j\omega})| = \frac{1}{|(1 - r_i e^{j\omega_{o_i}} e^{-j\omega})| |(1 - r_i e^{-j\omega_{o_i}} e^{-j\omega})|}, \quad (6.8)$$

For resonance, $|H_i(e^{j\omega})| \rightarrow \max$, therefore,

$$\frac{d|H_i(e^{j\omega})|}{d\omega} = 0, \quad (6.9)$$

solving the eq. (6.9) will give resonant frequency, ω_{r_i} ,

$$\omega_{r_i} = \cos^{-1} \left[\frac{1 + r_i^2}{2r_i} \cos \omega_{o_i} \right]. \quad (6.10)$$

considering pole radius, $r_i \rightarrow 1$ then we get,

$$\omega_{r_i} \approx \omega_{o_i}. \quad (6.11)$$

The impulse response of 2^{nd} order digital resonator is given by:

$$h_i[n] = Kr_i^n \sin \omega_{o_i} (n + 1) u[n], \quad (6.12)$$

where r_i is radius of poles, and K is the overall gain. The pole radius is *inversely* proportional to the -3 dB bandwidth. When radius = 1 (i.e., bandwidth = 0), sharp peaks in the spectrum are observed with the highest possible ($\sim \infty$) quality (Q)-factor. The change in pole radius corresponds to various energy losses, which is discussed in subsection 6.2.2. Due to the various energy losses, dissipation of energy occurs in the system, which causes the decrease in the amplitude of the resonances, leading to broadening of the -3 dB formant bandwidths. Thus, the effect of the damping factor r_i^n is observed. A relationship between -3 dB bandwidth B_i , and pole radius r_i for the i^{th} formant is given by [203],

$$r_i = e^{-\pi B_i T}, \quad (6.13)$$

where B is the -3 dB bandwidth (in Hz), and T is the sampling interval (in seconds). Therefore, for larger radius, sharp high peaks will be observed at the resonance frequencies. Hence, to achieve speaker anonymization, radius of the pole should be decreased so that there will be no presence of sharp and distinct peaks around formant frequencies, which will make identification of the speaker diffi-

cult.

6.2.2 Energy Losses

Ideally, the oral cavity is assumed to be a uniform tube with no losses due to the fact that the poles of corresponding transfer function equation (6.17) (which is a ratio of DTFT of volume velocities at lips and glottis) are strictly on $j\omega$ axis in s -plane. This oral cavity has roughly constant cross-section area with one end connected to the glottis and another at the lips [10,229]. However, in reality, this oral cavity can be modelled by time-varying and non-uniform cross-section area. The non-uniform tube model additionally considers the energy losses including vocal tract wall vibration, viscosity, thermal conduction of air particles, and the radiation loss at the lip, which are briefly described next.

- **Viscosity and Thermal Loss:** The effect of air particles in flowing from glottis to the lips have some friction with vocal tract walls, which resist the air-flow from the glottis. This friction can be introduced as a resistor in an electrical equivalent circuit of the cavity. (It should be noted that energy losses can be described by differential equations coupled to the wave equation that describes the pressure/volume velocity relations. However, these coupled equations are quite complicated, and it is difficult to obtain a closed-form solution to them. Therefore, the solution is found by a numerical simulation, requiring *discretization*, and hence a discrete element such as a resistor is introduced here.). This friction represents viscous energy loss. Another loss in the form of heat loss (also called as thermal loss) is incurred due to the vibrations of the vocal tract walls. A small decrease in formant frequencies and increase in formant bandwidth can be observed while considering these losses along with the wall vibration loss. The increase in bandwidth is more at higher frequencies [12].
- **Wall Vibrations:** Consider a tube whose cross-section is non-uniform. Furthermore, assume that the cross-sectional area changes slowly with time and space. The small differential sections of the surface of the wall ($d\Sigma$) are assumed to be independent by Portnoff [12]. He assumed the different pieces of the surface of the wall to be independent, i.e., *locally reacting*, because the change in cross-section due to pressure change is very small relative to the average cross-section. Each of these small sections, can be then mechanically modelled as shown in Fig.6.4, where $m_\omega =$ mass, $k_\omega =$ spring constant, and $b_\omega =$ damping constant per unit surface area.

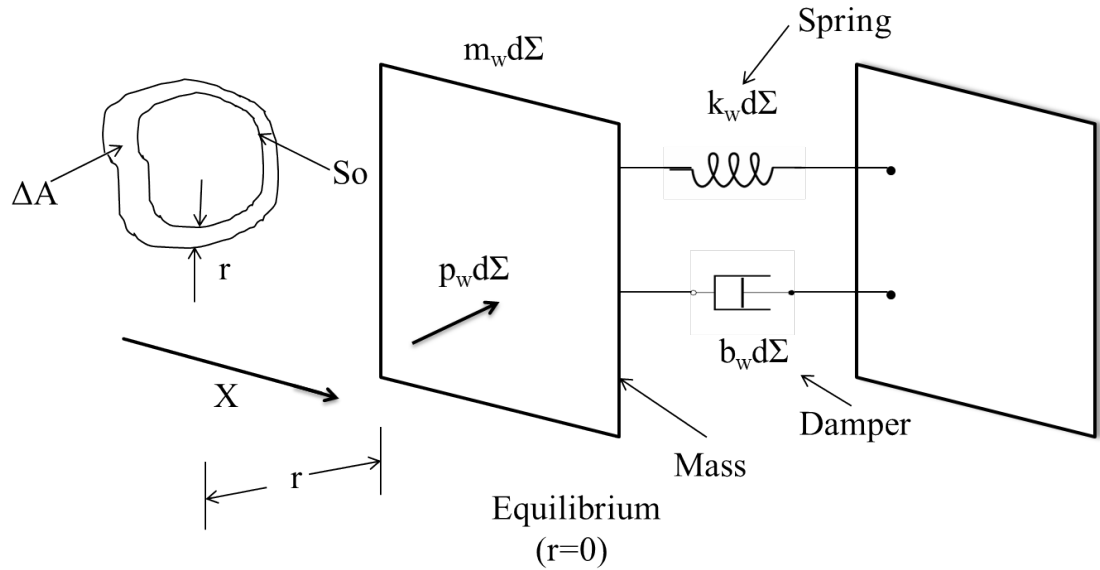


Figure 6.4: Mechanical model of differential surface element $d\Sigma$ of vibrating wall, after [10,12].

Considering the two boundary conditions of the volume velocity sources $u(0, t)$ (known), and the output pressure $p(l, t)$ (where $l =$ length of the vocal tract modeled as uniform tube), three coupling equations- two for sound wave propagation, and one 2^{nd} order differential equation from Figure 6.4 can be approximated as equation (6.14), (6.15), and (6.16), respectively [10].

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A_0} \frac{\partial u}{\partial t} \quad (6.14)$$

$$-\frac{\partial u}{\partial x} = \frac{A_0}{\rho c^2} \frac{\partial p}{\partial t} + \frac{\partial \Delta A}{\partial t}, \quad (6.15)$$

$$p = m_\omega \frac{d^2 \Delta A}{dt^2} + b_\omega \frac{d \Delta A}{dt} + k_w \Delta A, \quad (6.16)$$

where $A_0 =$ average cross-section(constant), $\Delta A =$ linear perturbation about the average cross-section, $S_0(x, t)$ is average vocal tract perimeter at equilibrium, r is perpendicular displacement of the wall, and $\rho =$ density of air particles. For the steady-state condition of the system described above, assume the system to be an LTI system. An input $u_g(t) = u(0, t) = U(\omega)e^{j\omega t}$ gives solutions, $p(x, t) = P(x, \omega)e^{j\omega t}$, $u(x, t) = U(\omega)e^{j\omega t}$, and $\Delta A(x, t) = \Delta \hat{A}(x, \omega)e^{j\omega t}$. Portnoff has used standard numerical simulation techniques to solve these coupled equations, which results in frequency response as

shown in Eq. 6.17 [12].

$$V_a(\omega) = \frac{U(l, \omega)}{U_g(\omega)}. \quad (6.17)$$

While producing voiced speech, due to air pressure from the lungs, glottis will vibrate (by invoking Bernouli's principle in fluid dynamics). Since vocal tract walls are *pliant*, they will move under pressure induced by sound propagation in the vocal tract system. These vibrations lead to energy losses in the cavity and hence, the poles of equation (6.17) are moved from the $j\omega$ axis, thereby becoming complex from being only imaginary (ideally). Hence, the -3 dB bandwidth is non-zero, and formant frequency is shifted. At low frequencies, inertial mass of vocal tract walls results in more motion, making it more dominant at lower frequencies compared to the higher frequencies [10].

- **Lip Radiation Loss:** The effect of radiation at lips can be analyzed by finding the acoustic impedance seen by the vocal tract from the lip end. This leads to the consideration of glottal and radiation load (at the lips) in the cavity model, as shown in Figure 6.5. R_r is the radiation resistance due to sound

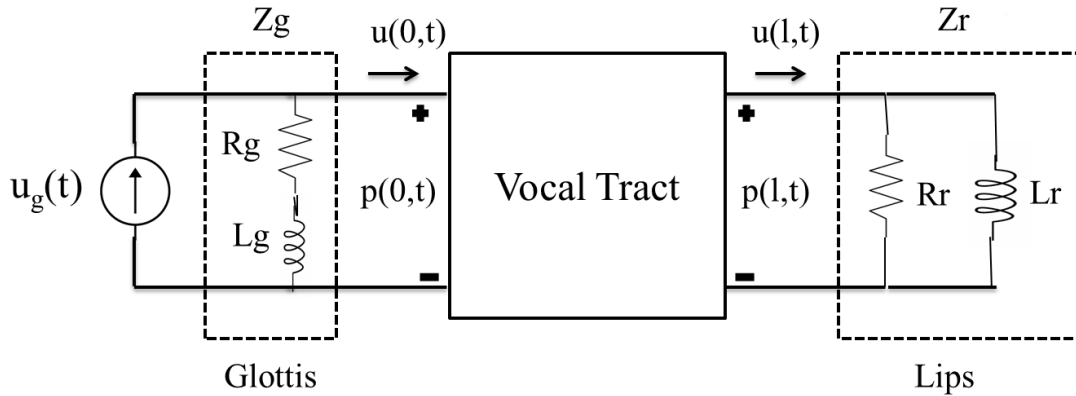


Figure 6.5: Glottal and lip boundary conditions as impedance loads. After [10].

propagation through lips and L_r is the radiation inductance, which is the inertial mass sent out at lips. Parallel combination of them contribute to the acoustic impedance as [10]:

$$Z_r(\omega) = \frac{P(l, \omega)}{U(l, \omega)} = \frac{1}{\frac{1}{R_r} + \frac{1}{j\omega L_r}} = \frac{j\omega L_r R_r}{R_r + j\omega L_r}. \quad (6.18)$$

For very small $\omega \approx 0$, $Z_r \approx 0$ so the radiation load acts as a short-circuit with pressure at the lips equal to zero, i.e., $p(l, t) = 0$. For very large ω with

condition $\omega L_r \gg R_r$, $Z_r \approx R_r$ making it resistive at higher frequencies. The radiation energy loss happens due to the real part of the complex impedance Z_r , which is proportional to R_r from eq. (6.18). Thus, more radiation loss will occur at higher frequencies with monotonic increase in R_r . From this discussion, it can be observed that this radiation impedance behaves as a Highpass Filter (HPF) [10]. Hence, to approximate the lip radiation, we can model the impedance as a HPF before we apply any algorithm on a speech signal.

Table 6.1: Frequency response of uniform tube with various losses with $p(l, 0) = 0$. After [10, 12].

Formants	Vibrating walls		Vibrating walls, viscous, and thermal loss		Vibrating walls, viscous, thermal, and radiation loss	
	Frequency (Hz)	Bandwidth (Hz)	Frequency (Hz)	Bandwidth (Hz)	Frequency (Hz)	Bandwidth (Hz)
1st	504.6	53.3	502.5	59.3	473.5	62.3
2nd	1512.3	40.8	1508.9	51.1	1423.6	80.5
3rd	2515.7	28.0	2511.2	41.1	2372.3	114.5
4th	3518.8	19.0	3513.5	34.5	3322.1	158.7

Table 6.1 shows the effect of energy losses on the formant locations and the -3 dB bandwidths. It can be observed that when all the losses are taken into account (i.e., vibrating walls, viscous, thermal, and radiation loss), there is a very high increase in -3 dB bandwidth for higher frequencies as shown in Table 6.1 [12]. Here, a comparison is made to the lossless system's formant values (i.e., odd multiples of 500 Hz) for a particular, case when tube length is 17.5 cm with the cross-sectional area of 5 cm^2 .

- **Relevance to the Design of Voice Privacy:** The most important thing to note here is that every human being has different configurations (in particular, size and shape) of the vocal tract system. In addition, the shape and size of the lips during speaking varies differently for everyone. These facts connect lip radiation loss to speaker-specific characteristics of a speech signal. As the speaker-specific characteristics lie in the higher formants (i.e., F_3 and F_4) [10], the energy losses become more important when we deal with the de-identification. In this chapter, we validate this hypothesis using various experiments which changes -3 dB bandwidth to change the speaker's identity.

6.2.3 Linear Prediction (LP) Model

In subsection 4.3.1 of Chapter 4, we described the LP model to predict a sample of speech using past p samples. However, in subsection 4.3.1, the focus was laid on

the amount of information carried by the LP residual and its dependence on the order p of the predictor, where the optimal order was found to be 8 for the replay SSD task on ASVSpooof 2019 PA dataset, having $f_s = 16$ kHz.

However, for designing a voice privacy system using linear prediction, the all-pole model is preferred because it is the acoustic tube model for speech production. The fact that Linear Prediction Coefficients (LPC) capture *implicitly* the time-varying area function of the vocal tract system, make LP model highly successful for various speech applications, more so, for speech coding. It can model sounds, such as vowels, well enough and the other consonants (except nasal consonants which requires zeros in the transfer function). The zeros arise only in the nasals and in the voiced sounds.

In LP analysis, sample at n^{th} instant is represented as a linear combination of past p samples, i.e.,

$$\tilde{s}(n) = \alpha_1 s(n-1) + \alpha_2 s(n-2) + \dots + \alpha_p s(n-p), \quad (6.19)$$

where $\alpha_1, \alpha_2, \dots, \alpha_p$ are called as LP coefficients. The z-domain system function for p^{th} order predictor is given as:

$$P(z) = \sum_{k=1}^p \alpha_k z^{-k}, \quad (6.20)$$

where p denotes the order of the predictor. The error signal or the LP residual signal, $e(n)$, is the difference between the actual (true) speech signal and the estimated speech signal. LP residual is given by:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^p \alpha_k s(n-k). \quad (6.21)$$

In z-domain, the error signal or LP residual $e(n)$ can be seen as the output of the prediction error filter $A(z)$ to the input speech signal $s(n)$, and is given by:

$$E(z) = A(z)S(z), \quad (6.22)$$

where prediction error filter $A(z)$ is defined as

$$A(z) = 1 - \sum_{k=1}^p \alpha_k z^{-k} = 1 - P(z). \quad (6.23)$$

The entire LP model can be viewed in two parts, namely, the analysis and the syn-

thesis. The LP analysis filter suppresses the formant structure of the speech signal and leaves a lower energy output prediction error, which is often called the LP residual (or LP error). LP residual is used as an excitation source for the production of speech, including its historical and commercially successful applications in speech coding [245]. The synthesis part takes the error signal as an input that gets filtered by the *inverse* filter, which is the inverse of the prediction error filter, and gives the speech signal as the output. When the vocal tract system is modeled as an LTI all-pole system, then a pole at $r_{o_i}e^{j\omega_{o_i}}$ and $r_{o_i}e^{-j\omega_{o_i}}$ correspond to i^{th} formant of the vocal tract system. Vocal tract length has *inverse* relationship with the formant frequencies. Thus, we can observe the difference in the formant frequencies between the male and the female speaker [246]. In particular, a male speaker (due to larger vocal tract length) tends to have lower formants than a female speaker [10]. It should also be noted that for the design of voice privacy system, the LP order is taken to be 16 (10 for vocal tract, 4 for lip radiation, and 2 for glottal flow), unlike the case of LFRCC feature set (where optimal LP order was 8) in Chapter 4, where the goal was replay SSD.

- **Relevance to Speaker Anonymization:** In LP model, LP coefficients govern the pole locations, which in turn determine formant frequency and formant bandwidth [247]. Mathematically, formant frequency is given by $\frac{F_s\theta}{2\pi}$, where θ is the pole angle in radians, given F_s is the sampling frequency in Hz. The formant bandwidth is given by $\frac{F_s}{\pi}(-\log(r))$, where r is the radius of the pole [10]. As per M.R. Schroeder, "human beings emit and perceive sounds by emitting spectral peaks more dominantly than the spectral valleys" [248]. Therefore, we can achieve speaker de-identification by modifying the formant frequencies leading to the change in the formant spectrum with *naturalness* and *intelligibility* retained in the anonymized speech. Hence, by performing controlled shift in the pole angle and the pole radius, speaker de-identification can be achieved without the loss of intelligibility in the anonymized speech signal. However, it should be noted that *significant* modifications to the formants will definitely affect the intelligibility and the naturalness. The modifications done in our work are *NOT significant* to hamper the naturalness and intelligibility, and are *just enough* to have altered speaker's identity. The impact of these modifications is evaluated by metrics such as EER (measuring the impact on speaker's identity) and WER (measuring the impact on intelligibility).

Algorithm 9 Voice Privacy by LP Modelling of Speech Production

- 1: **procedure** VOICE_PRIVACY(x) ▷ x is the speech signal
 - 2: LP coefficients and residuals are extracted.
 - 3: LP coefficients are converted to poles.
 - 4: Radius of the complex poles is shifted to 0.975 of the original value of radius.
 - 5: Poles ϕ of the complex poles are shifted to $\phi^{0.8}$.
 - 6: New LP coefficients are formed.
 - 7: The new anonymized speech signal is re-synthesized.
 - 8: **end procedure**
-

6.2.4 Proposed Voice Privacy System

At first, the speech signal is divided into frames of 30 ms duration, with an overlap of 15 ms, which are fed to LP source-filter analysis in order to obtain the LP coefficients and residual. Only LP coefficients are taken into account for further processing, while the residual is left unchanged (as compared to its use in LFRCC for replay SSD in Chapter 4). LP coefficients are then employed to obtain the pole positions of the LP model. The poles, whose imaginary value is not zero, are considered, and their pole angle ϕ is calculated. Since most of the complex conjugate pole-pair correspond to one formant frequency each, only one pole out of the complex conjugate pole-pair is considered for achieving speaker anonymization [249]. To further improve the baseline system, the pole radius is changed along with the pole angles. The pole angle is shifted by raising it to the power of McAdams coefficient $\alpha = 0.8$, i.e., ϕ^α [250]. Values of α and ϕ determines the positive or negative shift in the pole locations. The pole radius is decreased by 15%, 5%, and 2.5% of original pole radius [15]. With this new set of pole radius and angles, new set of poles are fabricated, and therefore, forming new LP coefficients. These new coefficients along with original LP residuals are used to synthesize new speech signal and hence, achieving the anonymization of speech. The functional block diagram of the same is as shown in Figure 6.6. Motivated by original studies in speech coding literature [251–253], residuals are kept intact because they are used to retain the naturalness and intelligibility of the speech signal.

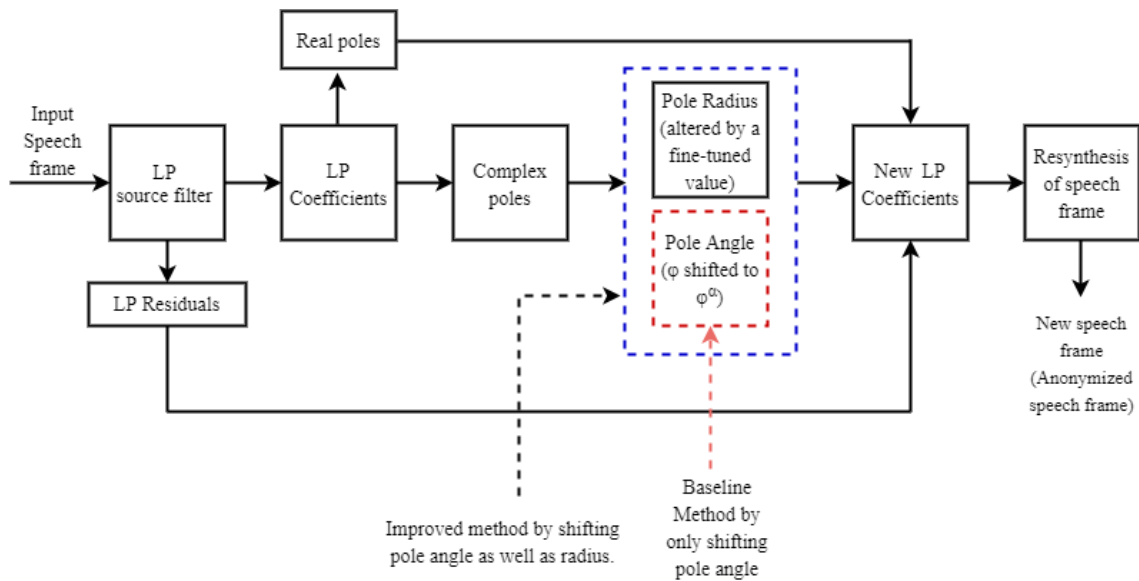


Figure 6.6: Proposed LP-based anonymization system. After [13–15].

6.3 Experimental Setup

6.3.0.1 Datasets Used

For development data, subsets from two corpora, namely, LibriSpeech-dev-clean and VCTK are provided [254, 255]. These subsets are further divided into trial and enrollment subsets. There are 40 speakers in LibriSpeech-dev-clean. There are 29 speakers in enrollment utterances, and 40 speakers in trial utterances. From these 40 speakers of trial subset, 29 speakers are also included in the enrollment subset. In VCTK-dev dataset, there are total 30 speakers, which are the same for both trial and enrollment utterances. Furthermore, for trial utterances, there are two parts, denoted as *common part* and *different part*. Both the parts are disjoint in terms of utterances, however, they have the same set of speakers. The *common part* of the trials has utterances from #1 to #24 in the VCTK corpus, which are the same for all the speakers. The *common part* of the trials is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. #25 onward utterances are distinct and hence, are included in the *different part* of the VCTK-dev dataset. For evaluation, the structure is the same as that of Dev set, except for the number of utterances.

6.3.1 Experimental Results

In the experiments, decreasing the radius of the poles (i.e., r) results in the expansion of the formant bandwidth B . On studying the experimental results, it is

Table 6.2: Statistics of the Dev Datasets. After [13].

Subsets of corpus	Particulars	Female	Male	Total
Librispeech: Dev-clean	Speakers in enrollment	15	14	29
	Speakers in trials	20	20	40
	Enrollment utterances	167	176	343
	Trial utterances	1018	960	1978
VCTK-dev	Speakers (same in enrollment and trials)	15	15	30
	Enrollment utterances	300	300	600
	Trial utterances (common part)	344	351	695
	Trial utterances (different part)	5422	5255	10677

Table 6.3: Statistics of the Eval Datasets. After [13].

Subsets of corpus	Particulars	Female	Male	Total
Librispeech: test-clean	Speakers in enrollment	16	13	29
	Speakers in trials	20	20	40
	Enrollment utterances	254	184	438
	Trial utterances	734	762	1496
VCTK-test	Speakers (same in enrollment and trials)	15	15	30
	Enrollment utterances	300	300	600
	Trial utterances (common part)	346	354	700
	Trial utterances (different part)	5328	5420	10748

observed that 2.5% decrease in the radius along with the phase changed to $\phi^{\alpha=0.8}$, gave higher values of %EER, and lower values of %WER, which is desired. As discussed earlier, by decreasing the pole radius, corresponding formant bandwidth will increase. According to the digital 2^{nd} order resonator (filter) theory (discussed in subsection 6.2.1), an increase in the bandwidth will decrease the quality factor Q of the resonator. Spectrum peaks will no longer be distinctly present, causing the loss of speaker-specific information. Hence, the quality of original speech signal degrades, which in turn contributes to the higher EER scores. The results of the experiment in terms of %EER and %WER for test data and Dev data are shown in Figure 6.7a, 6.7b, 6.7c, 6.7d [15, 256].

6.3.1.1 Gender-Based Analysis

From the experimental results obtained for voice privacy, it can be observed that the %EER values are higher for the female speakers than the male speakers under the condition that the anonymization technique on the utterances is the same for both the female and the male speakers, as shown in Figure 6.7a and Figure 6.7b. This result can be supported by the fact that spectral resolution for female speech is poor as compared to male speech [16]. The mass of the vocal folds in female speakers is less than the male speakers due to which movement of vocal folds becomes sluggish in male speakers hence, the glottal vibrations are more rapid (faster) in female speakers, and therefore, high pitch frequency is observed for female voice. Hence, in the spectral-domain, the pitch source harmonics are

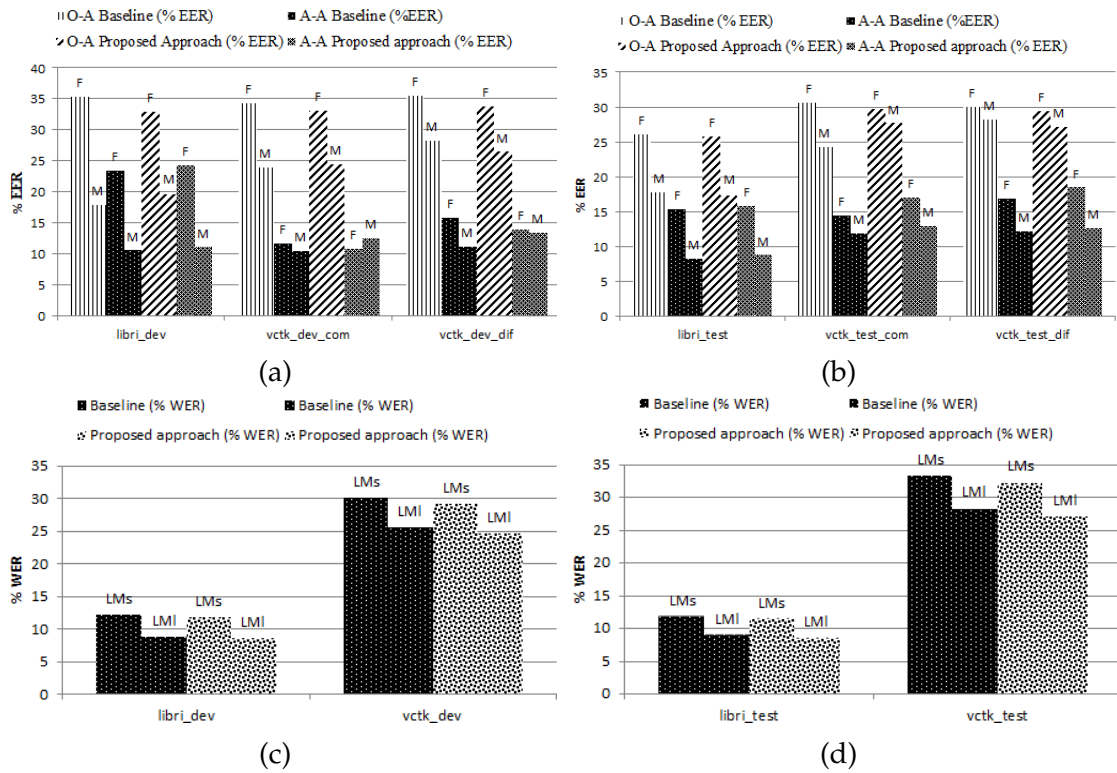


Figure 6.7: %EER (o- original, a- anonymized) for (a) Dev data, (b) test data, and %WER (for two trigram LMs: LM_5 -small LM, and LM_1 -large LM) for (c) development data, (d) test data (for radius = 0.975 to its value and $\alpha = 0.8$).

observed to be at a greater distance with each other, which results in the poor spectral resolution of the female speaker. This can be beneficial for our aim to achieve speaker de-identification because the recognition of female speakers through ASV systems can become difficult. In addition, in the glottal cycle waveform, glottal closure instant, and period during closure provide characteristics for discriminating speakers' voices from one another. Provided the same pitch period and impulse response of the vocal tract system, even a slight variation in the glottal flow waveform can result in a considerable amount of change in the voice characteristics. Therefore, due to the larger duration of pitch periods in male speakers, they get sufficient time for the closure of the glottis and to perform activity (i.e., nonlinearities introduced near the (Glottal Closure Instances) GCIs) near the glottal closure. However, in the case of female speakers, the pitch period duration is almost half the pitch period of the male speakers (near about 10 ms in male speakers and 5 ms in female speakers), due to which female speakers don't have enough time between the closures as compared to male speakers [257]. This large variation in the glottal waveform changes the speaker's characteristics drastically, and hence, the difference in male and female speech. The speaker recognition tech-

niques use information based on the 1-2 ms glottal closure period [258]. Hence, tracking this large variation in 1-2 ms of glottal closure period becomes difficult for the ASV systems, which can lead to higher %EER values.

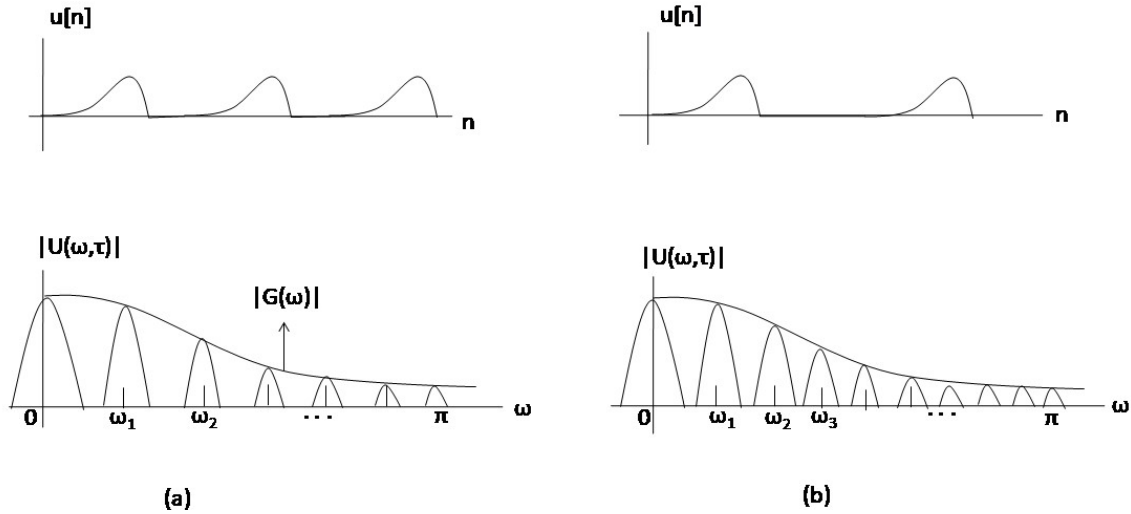


Figure 6.8: Illustration of periodic glottal flow and its spectrum for (a) higher pitch of female speaker, and (b) lower pitch of male speaker. After [10, 16].

The illustration of the spectral resolution problem is presented in Figure 6.8. In particular, $u[n]$ represents the glottal flow waveform model which can be given by,

$$u[n] = g[n] * p[n], \quad (6.24)$$

where $g[n]$ is the glottal flow waveform over a single glottal cycle, and $p[n]$ is an impulse-train [10]. $U(\omega)$ and $G(\omega)$ is the Fourier transform of $u[n]$ and $g[n]$. ω_k denotes the harmonics of the glottal flow waveform. The magnitude of the spectral shaping function, $G(\omega)$ is referred to as the spectral envelope of the harmonics.

Furthermore, the analysis of the spectrogram of original and anonymized speech of both the female and male speakers is shown in Figure 6.9. The original speech of both the male and female speakers has undergone the same anonymization method, which was discussed in subsection 6.2.4. According to the change in the pole angle ϕ , corresponding formants will be shifted, i.e., for $\phi < 1$ the formants will be shifted to a higher value and vice-versa. Due to this reason, lower formants in male speech will shift to a higher value and high pitch-source harmonics for female speech will be observed.

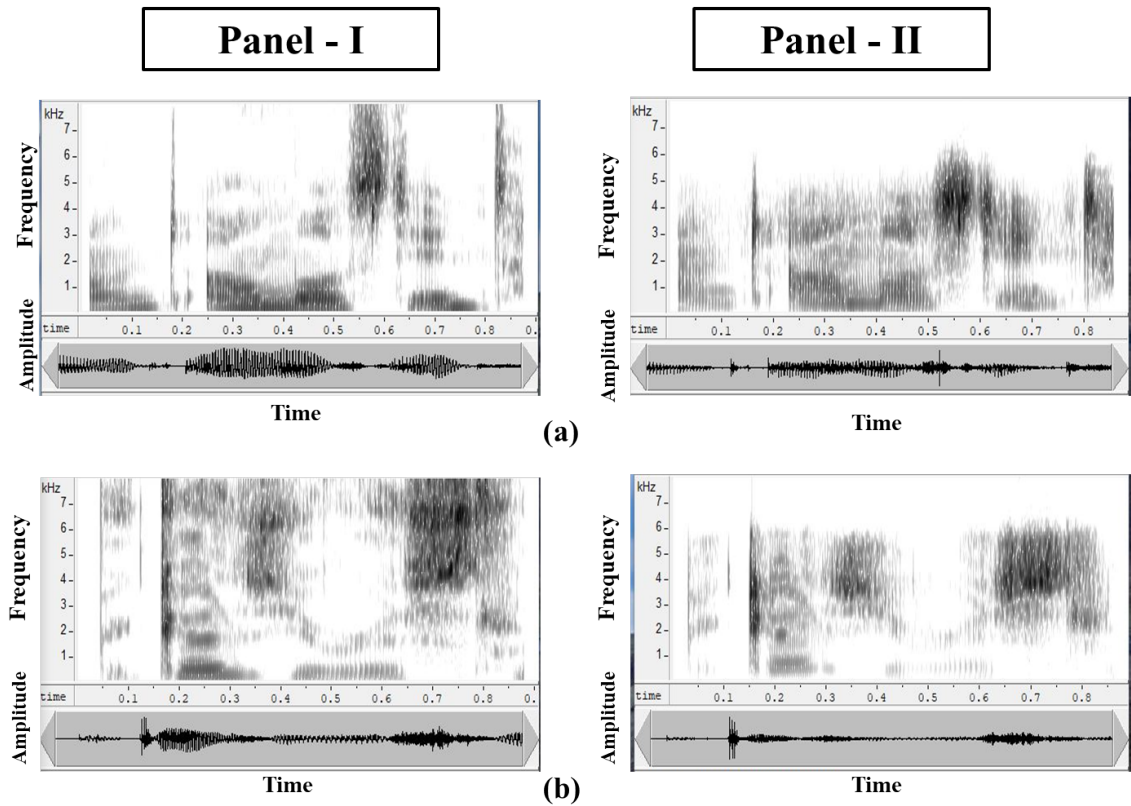


Figure 6.9: Panel-I : Analysis of original speech signal. Panel-II: Analysis of anonymized speech signal: (a) spectrogram and speech signal for a female speaker, and (b) spectrogram and speech signal for a male speaker.

6.4 Voice Privacy and Attacker’s Perspective

This section first discusses the attacker’s approach to target selection, followed by the discussion on how voice privacy can help in misleading the attacker. To that effect, understanding the attacker’s perspective is also important. Hence, we present the approach of target selection by the attacker in the following subsection.

6.4.1 Target Selection

In this section, we discuss one of the possible vulnerabilities as target selection. In order to increase the chances of a successful attack, an attacker selects the most vulnerable target by using the attacker’s own ASV, as shown in scenario 2 of Figure 6.10 [259]. The attacker chooses the most vulnerable target as the speaker contributing the most to the FAR. This strategy of selecting the most appropriate (i.e., most vulnerable) speaker is known as *target selection*. Furthermore, this approach of target selection is primarily focused on mimicry attacks on speakers, particu-

larly those speakers whose speech data is available in public in large quantities, such as celebrities. With the aim to protect the privacy of speakers, the approach of target selection can also be useful in determining how secure a closed-domain targeted ASV system is [260].

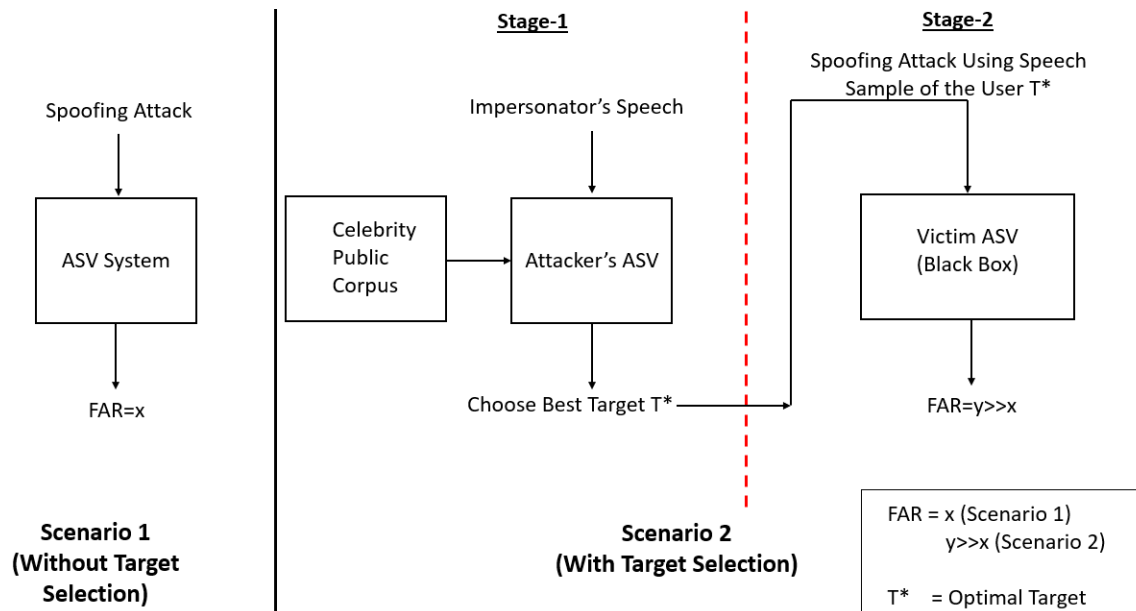


Figure 6.10: Target Selection: By using the Attacker's ASV to Attack the Victim's ASV.

To further intuitively understand the attacker's approach of target selection, we assume Log-Likelihood Ratio (LLR) as being the *similarity score*. Usually, it is compared to a predefined threshold, which then defines the FAR and FRR. Additionally, the LLR is computed in terms of the Probabilistic Linear Discriminant Analysis (PLDA) score for the state-of-the-art x-vector based approach of ASV [261]. Target selection attacks can be performed in one of the following two ways:

1. By selecting the most vulnerable speaker (from the speaker classification as shown in Figure 6.11), referred to as '*lamb*' in [259], from the set of enrolled speakers. Lambs are the speakers, who are easiest to mimic w.r.t. a specific attack. Thus, the speaker with the highest LLR score w.r.t. that attack, is selected as being the lamb.
2. By selecting the most skillful attacker (i.e., the speaker who is successful at imitating other speakers), referred to as '*wolf*' in [259], w.r.t. a pre-defined victim speaker. Thus, an attacker with the highest LLR score w.r.t. the fixed pre-defined victim, is selected as being the wolf.

It should be noted that the target (i.e., the most vulnerable speaker) is also dependent on the type of attack, or the algorithm used for spoofed speech generation. More so, if the attack is a mimicry attack, the selection of the target also depends on the attacker's speech. The dependency for each of the spoofing attacks is discussed as:

- In the case of spoofing by professional mimics, suprasegmental or prosodic features of the target speaker are used for imitation [1]. Furthermore, the use of these prosodic features is dependent on the relative skillfulness of the mimics [257].
- In the case of twins, spoofing is indeed dependent on the speaker type, because the co-twins share similar segmental information in their speech signals.
- For the case of spoofing attacks such as replay, the success of the attack depends on the quality of recording and playback devices, and also the environmental or acoustic conditions (for example, more reverberation in a room leads to the attack being easily detected by the SSD system).
- For the case of spoofed speech generated due to voice conversion, the algorithms for voice conversion convert the speech of a source speaker into the speech of a target speaker. Hence, the attack is speaker-dependent in this case.

Notably, target selection is different from a speaker identification perspective. In the latter, the claimed identity is compared with all the speaker models, and the maximum closeness to the claimed identity is compared with a pre-defined threshold. Contrary to this, in target selection, as shown in Figure 6.10, no single speaker is claiming his/her identity and hence, the ASV system has to be run iteratively to include all the speakers. Moreover, the chosen target is responsible for the maximum FAR, out of all the speakers.

6.4.2 Target Selection by the Attacker and Voice Privacy System

According to Doddington's menagerie [259], a pool of speakers can be divided into 4 categories, based on their effect on the EER, as shown in Figure 6.11. Most of the population of the speakers are of *sheep* type, which is the default speaker type, for which the ASV system performs well. However, target selection is performed on the remaining speakers (i.e., goats, lambs, and wolves type), which contribute

to the EER (i.e, FAR or FRR). It should be noted that only 15 to 20% of the speakers contribute to %EER [259]. Since each of the type of speakers has a different effect on the EER, they can also be classified in terms of their vulnerability levels to attacks, as shown in Figure 6.11. From an attacker’s perspective, the *lamb* type

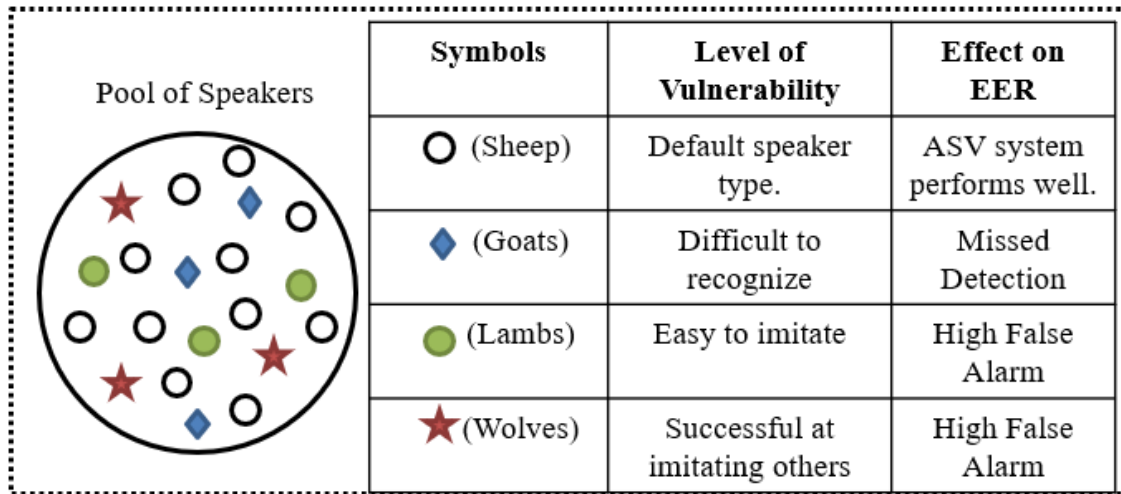


Figure 6.11: Types of speakers based on their vulnerability levels and their effect on EER scores.

speakers should be the target. The attacker can perform target selection using an ASV system, to choose the most vulnerable speaker from the pool of speakers, as the target [2]. However, if voice privacy is used, the target selection procedure by the attacker will yield incorrect results. Thus, the attacker will be fooled into selecting a not-so-vulnerable speaker as the target, as shown in Figure 6.12.

If the anonymization is achieved on the corpus by using a voice privacy system, the actual speakers are mapped to their corresponding *pseudo-speakers*. The output of a voice privacy system is a speech utterance, which sounds as if it had been spoken by another speaker, known as *pseudo-speaker*. Thus, the anonymization process alters the identity of the speaker, however, retains its intelligibility and naturalness [13]. The change in identity modifies the vulnerability levels of the speakers. Therefore, as shown in Fig. 6.13(b), after voice privacy is applied, the target selection system yields pseudo-speaker C as the most vulnerable one, whereas in reality, speaker A is the most vulnerable. Hence, the attacker can be fooled into believing that the target is pseudo-speaker C. Therefore, the technological challenge faced by the attacker is due to the anonymization provided by the voice privacy system on a speech corpus.

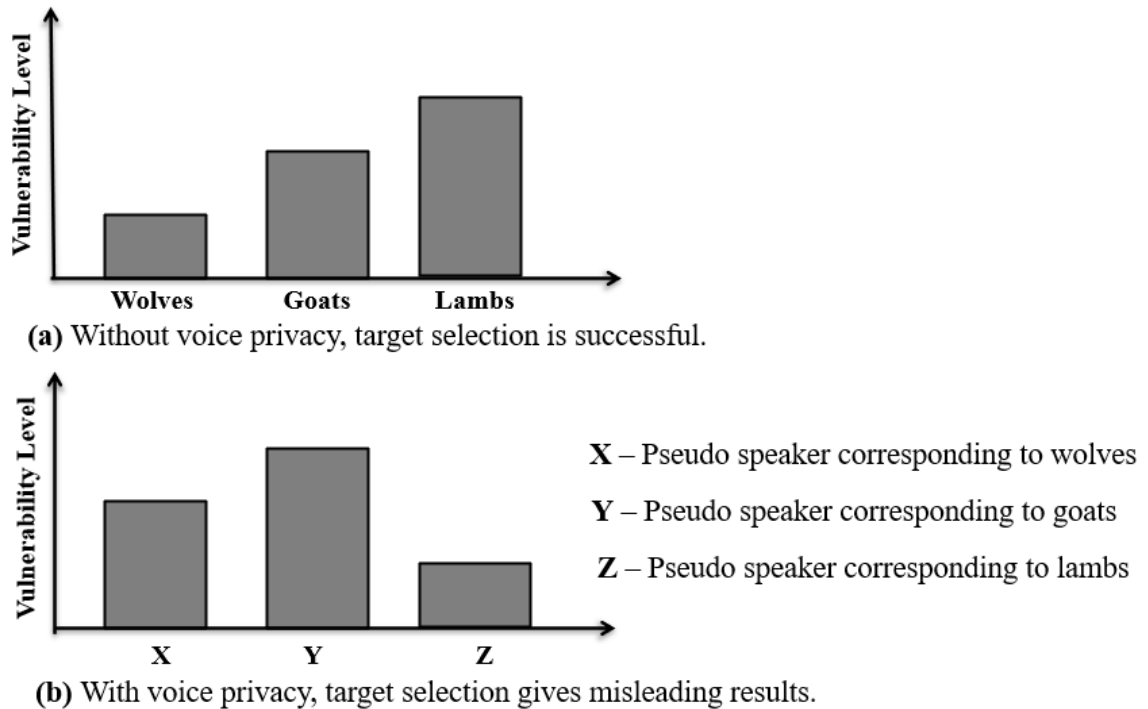


Figure 6.12: Schematic representing effect of Voice Privacy on target selection.

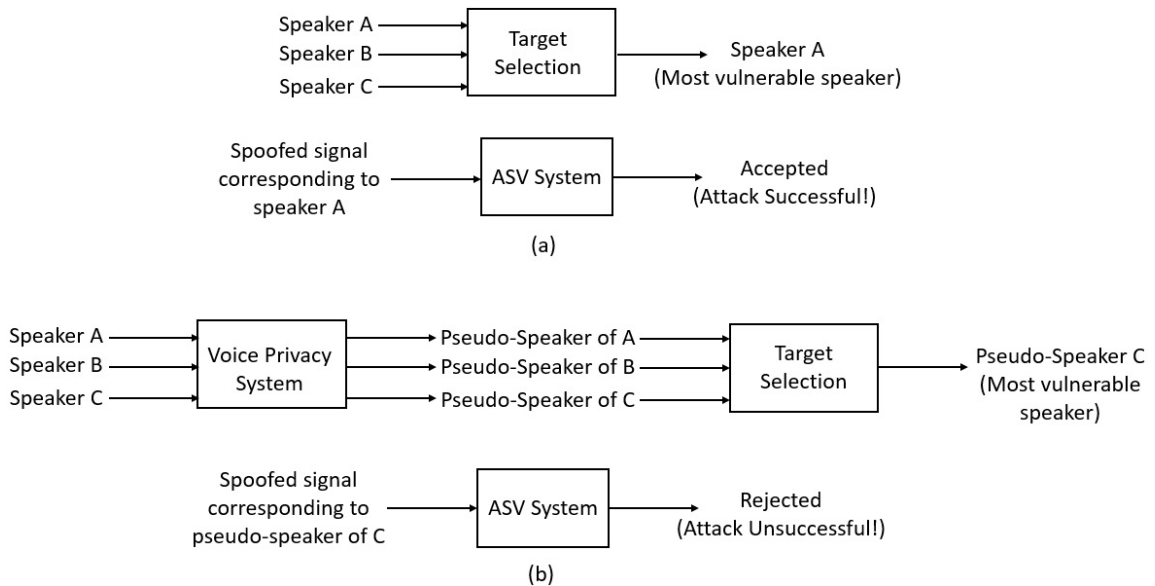


Figure 6.13: (a) Attack using target selection, but without voice privacy system, and (b) attack using target selection with voice privacy system.

6.5 Target Selection in Enrolled Users with Malicious Intent

In principle, an enrolled user has more power to attack than an attacker (usually a non-enrolled outside entity). An enrolled user with malicious intent may attempt

to spoof the system, which is, all the more, a greater security concern/threat. A real-world example of this type of attack is the twin fraud in HSBC bank, where the bank's voice authentication system was spoofed by a BBC journalist, and his non-identical co-twin speaker [44], [45]. Another interesting point to note here is that if an SSD system as a countermeasure for twins attack is used, it will prevent malicious twins from impersonating (which is based on physiological characteristics, in particular, the shape and size of the vocal tract system [1]). However, it will also prevent genuine, and zero-effort imposters from verification and hence, increasing the False Rejection Rate (FRR). With the deployment of a Voice Privacy system instead, this kind of attack will not be possible. Moreover, the issue of preventing genuine and zero-effort imposters will also be alleviated and hence, there will be no increase in FRR.

The work reported in [257] is based on 17 pairs of twins. We performed an experimental analysis under the scenario when one co-twin speaker attempts to mimic its other co-twin counterpart. To that effect, we investigate the twin-pair that has the most similar co-twin, based on the EER outcome found iteratively for each twin-pair.

6.5.1 Setup

- **Dataset Used:** The twins data used in this work has been taken from [257], with sampling frequency as 22050 Hz. The Hindi (an Indian language) subset of the dataset was used. It was prepared with the help of tape recorders (Sanyo model no. *M – 1110C* and Aiwa model no. *JS299*) with microphone input and close-talking microphones (namely, Frontech and Intex). The total number of pairs of twins was taken to be 17. The speakers belonged to various dialectal regions of Maharashtra (a state in India), with a native language as Marathi (an Indian language). Out of the 17 twin-pairs, 12 pairs of twins are male-male siblings, 4 pairs of twins are female-female siblings, and 1 pair of twins is male-female siblings. The speakers' age in the dataset varies from 7 years to 61 years, at the time of recording of data.
- **Data Preparation, Features, and Classifier Used:** The available twins corpus has speech data corresponding to 17 twin-pairs. For training data, each of the recordings of each twin-pair was of about three to five minutes duration. For evaluation data, each of the recordings of each twin-pair was of about one-to-two minutes duration. Therefore, each speech recording was divided into chunks of 5 sec of data. This was done for both the training and

evaluation datasets.

- **Data Augmentation used:** To overcome the limitation of limited data, we have augmented the data for better training of the classifier. The data augmentation is done using a concept of phase shift, as also done in [262]. The human auditory perception is not affected, when the phase of the speech signal is shifted by 180 degrees [212]. Therefore, we shift the phase of our speech recordings by 180 to obtain another set of speech data. It is important to note that this data augmentation strategy does not affect the speaker's identity.
- **Feature and classifier used:** The state-of-the-art MFCC feature set was used with dimension 40. The classifier used was based on Gaussian Mixture Model (GMM) with the number of Gaussian components kept as 128, considering resource constraints.

6.5.2 Experimental Results

The most vulnerable twin pair is selected on the basis of %EER. Figure 6.14 shows the EER obtained for each twin-pair. It can be observed that the twin-pair 6 (con-

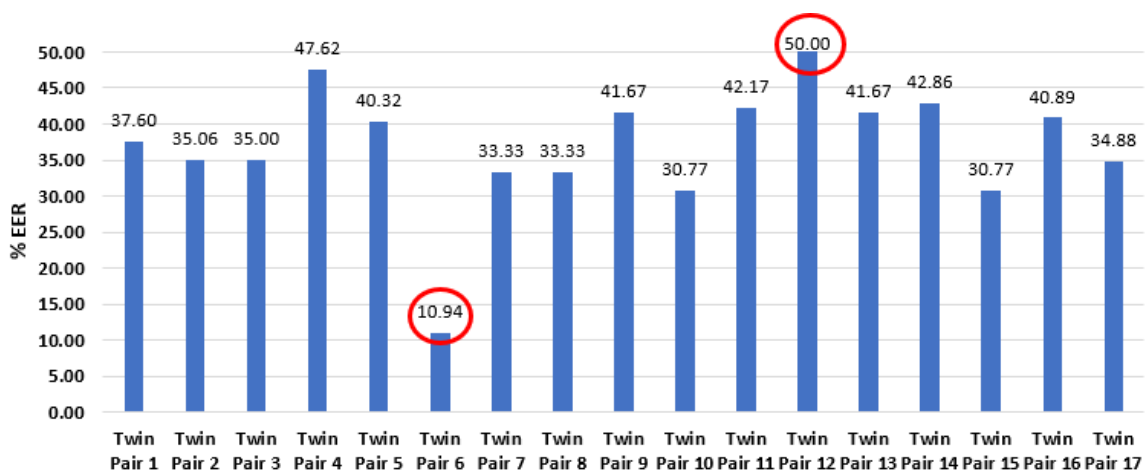


Figure 6.14: A case study on target selection: EER estimation of each twin-pair.

sisting of male-male speakers) has the lowest EER of 10.94%. It indicates that the twin-pair 6 consists of the most similar speakers. According to Dodding-ton's menagerie, twin-pair 6 is of *lamb* type speaker, as twin-pair 6 leads to the least EER, and hence, the co-twins of this twin-pair can mimic each other with the highest probability of success, as compared to the remaining twin-pairs in the dataset. Furthermore, the twin-pair 12 (consisting of male-male speakers) has the

highest EER of 50%, indicating that the co-twins are the most dissimilar speakers and hence, twin-pair 12 is the least vulnerable to twins attack.

Out of these 17 twin-pairs, 4 pairs are of female-female type, 12 pairs are of male-male type, and 1 pair is of male-female type. The average EER of female-female twin-pairs is 10%, and the average EER of male-male twin-pairs is 26.08%. The effect of gender on vulnerability remains an open research direction.

6.6 Technological Challenges Faced By the Attacker

In this section, we study and analyze various issues the attacker faces in order to attack any given ASV system [2].

6.6.1 Number of Trials on Victim ASV Access

In realistic scenarios, an effective ASV system should have an upper limit to the number of trials that can be allowed for a particular speaker. However, an assumption for target selection attacks is that the attacker can have in principle, an infinite number of trials (since the attacker uses his/her own ASV to attack), in order to effectively practice the mimicry, which is impossible in practical scenarios of ASV system development.

6.6.2 Corpora for Attacker's Perspective

The attacker can proceed with the target selection attacks only when the corpora used for ASV is public, such as VoxCeleb. This is because target selection should be performed over the same corpora as that of the victim ASV. If this is not the case, then the probability of a good LLR score will decrease drastically, as the probability of the existence of a speaker, who is also the most vulnerable in two different datasets is almost negligible. Thus, the system becomes error-prone.

Moreover, there are various corpora available in the literature w.r.t. anti-spoofing research, such as ASVSpooF 2015, 2017, 2019, and 2021 datasets. However, these standard datasets are limited to a fixed number of configurations of data collection setup, and thus, limited recording conditions. Moreover, datasets are prepared with certain underlying assumptions. Such assumptions keep us far away from developing generalized anti-spoofing systems suitable for real-world applications. For example, the generation of spoof utterances in the ASVSpooF 2015 dataset is limited to 10 algorithms of VC and SS. Similarly, the replay spoofing utterances in the ASVSpooF 2017, 2019, and 2021 datasets are limited to a fixed

number of recording configurations. This makes the attacker mount complementary attacks by utilizing the weakness of the underlying SSD system because till now the corpora for anti-spoofing are limited to a specific attack only. Therefore, we are far away from designing a versatile SSD system, which would alleviate all the five types of presentation attacks as well as unknown attacks. Additionally, these publicly available corpora are available to the attacker as well. To that effect, attacks over unprotected corpora can be used to determine personal information about speakers using techniques, such as *target selection*, which enables an attacker to select the most vulnerable speaker from a corpus [2,260]. Figure 6.15 shows a Venn diagram w.r.t. the publicly available corpora for developing anti-spoofing defences against various spoofing attacks. However, it should be noted that there exists no dataset (this situation is denoted by '?' in Figure 6.15), which aims at developing CMs for more than one or *all* the spoofing attacks. Therefore, there is still a long way to come up with generalized CMs, that are suitable for real-world SSD deployment.

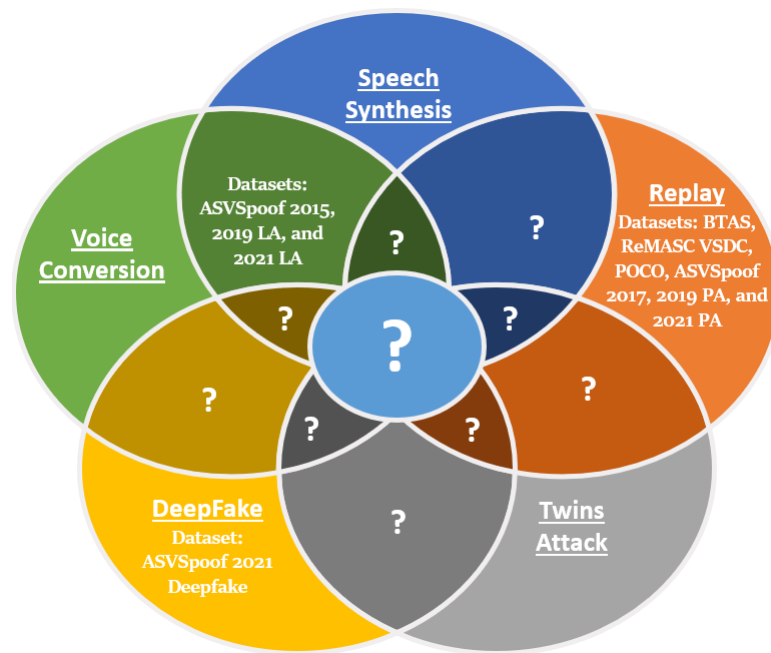


Figure 6.15: Publicly available corpora for anti-spoofing research and the associated known attacks. Here, '?' indicates, a gap area to develop anti-spoofing corpora from attacker's perspective.

6.6.3 Transmission Channel

As per the recent anti-spoofing literature, transmission channel conditions are known to play an important role in the performance of SSD systems. Hence, anti-

spoofing over a phone channel was chosen as the topic of the recent ASVSpooF 2021 challenge [263]. Thus, the transmission channel also forms one of the technological challenges from the attacker’s perspective as well.

6.6.4 Perturbation Minuteness in Adversarial Attacks

While attacking by adversarial ML approach, the boon for the attacker can even become disadvantageous. The small perturbation might not be captured over the air, causing the attack to be unsuccessful, especially in the case of Voice Assistant systems [264]. Consequently, over the air, the performance of perturbed signals should also be considered, while evaluating the chances of a successful attack by adversarial ML methods. Furthermore, the perturbation should be such that it bypasses any smoothing technique used in the ASV system [265].

6.6.5 Voice Privacy Systems

As per the discussions in Section 6.1 and Section 6.2, Voice Privacy (VP) aims to hide a speaker’s identity, retaining the speech linguistic content and naturalness [13,114]. If the users publish data without anonymization, the attacker gains illegal access to it, and can further use speakers’ information to attack the ASV system (as shown in Figure 6.16). If a speech signal undergoes a considerably good algorithm for VP and anonymized data is published, it will be almost impossible for an attacker to perform target selection due to the absence of mapping between the speech data and actual speaker identity [15,266]. If VP is used, the most vulnerable target T^* cannot be chosen correctly. Consequently, the approach of *optimal* target selection will not be useful, and the attacker will be left with only fewer attack strategies.

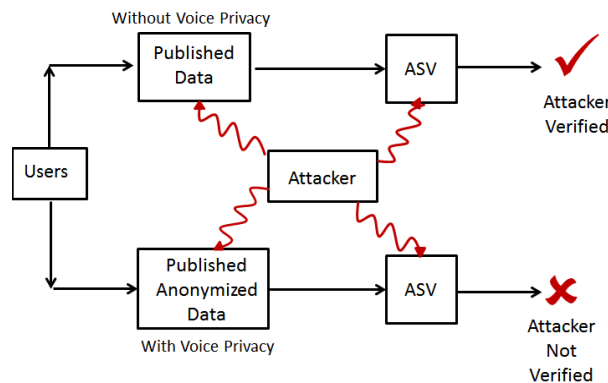


Figure 6.16: Game between an attacker and VP system.

6.6.6 Voice Liveness Detection (VLD)

The countermeasure solutions developed in the ASVSpooF Challenges are specific to a particular attack. Given that the attacks on the ASV systems can be known or unknown attacks, VLD systems aim to detect only the live speech signal, and reject all the other *non-live* speech, which are generated from known and unknown attacks [26]. In this context, as discussed in Chapter 5, VLD is an emerging research area in which pop noise has been used actively as a discriminative acoustic cue to detect live speech [92, 102, 267]. VLD systems enhance the security of the ASV system. Given that VLD systems aid in enhancing the robustness against attacks on ASV, they have also become a technological challenge for attackers. In particular, VLD systems are highly efficient against replay attacks. Replay attack requires only a recording device to capture a genuine user’s voice from a distance. However, due to the distance of the recording device from the speaker, liveness cues, such as pop noise are faintly captured, or even absent in some cases. Moreover, even in the case of artificially synthesized signals, a playback device/loudspeaker is needed to mount the attack, which in turn diminishes the strength of pop noise, which is strongly present in live speech. Moreover, till now, VLD is analyzed w.r.t. only replay attacks. However, the scope of VLD in other spoofing techniques, such as VC and SS, remains to be explored and is an open research problem.

6.6.7 DeepFake Detectors

Advances in DeepFake generation techniques have made fake data each time more accessible. Thus, DeepFake detection has gathered immense interest, especially in images and videos [268, 269]. Nevertheless, we focus our discussion on speech DeepFake detectors, which have not been considered as much as image and video DeepFake detectors. In [270], higher-order power spectrum correlations are considered in the frequency-domain. Bi-spectral characteristics, such as bi-coherent magnitude and phase spectra, were used to observe third-order correlations. Differences were observed in the bi-coherent magnitude and phase spectra between natural and synthetic speech. In [271], semantically rich information was extracted by using latent representation. Particularly, *XcepTemporal* convolutional recurrent neural network was introduced for DeepFake detection by stacking multiple convolution modules.

6.7 Voice Privacy and Cryptography

Cryptography aims to prevent any malicious usage of data by *encrypting* data to an unreadable (or unrecognizable) form. There are two primary types of cryptographic algorithms - symmetric key (private key) encryption and asymmetric key (public key) encryption. The symmetric key encryption deals with a *single private key*, which is used for encryption *as well as* decryption. Therefore, the symmetric key encryption itself can require security for the protection of the key. Moreover, the total number of keys required for p parties should be $p(p - 1)/2$ [17]. Hence, due to these key-management issues, public key encryption has taken over most of the security applications. Unlike the symmetric key encryption, in the public key encryption (i.e., asymmetric key encryption), encryption and decryption are performed using *two different keys*- one of them is a public key, which is used for encryption, the other is a private key, which is used for decryption [272]. The Rivest Shamir Adleman (RSA) algorithm is one of the most standard algorithms used for private key encryption [273]. In the following subsections, we will discuss the most widely used RSA algorithm and its time complexity.

6.7.1 Public Key Encryption

Public key encryption uses two kinds of keys - public and private. Public keys are accessible to everyone, including the attacker. However, private keys are known only to a single user. Each user has his/her own private key, which is not to be shared with anyone. As shown in Figure 6.17, the message is encrypted by the

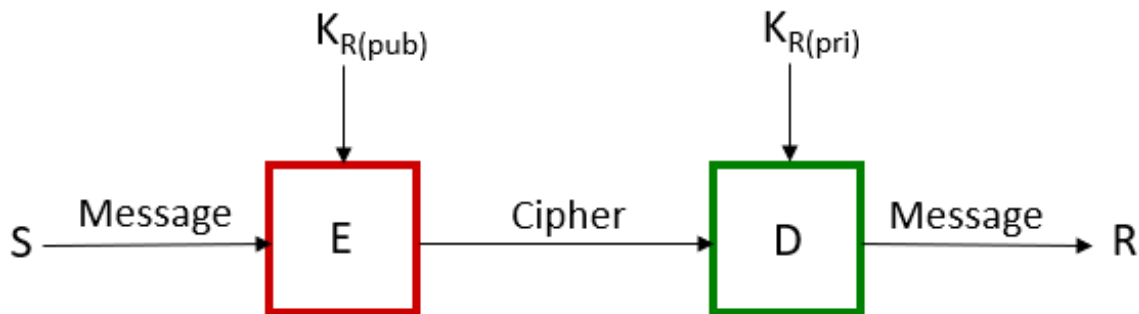


Figure 6.17: Public key Encryption and Decryption. After [17].

sender S , with the help of the receiver's public key, $K_{R(pub)}$. Examples of public key algorithms are RSA, Diffie-Hellman, and El-Gamal encryption [17].

Though key management is not a major issue with such types of algorithms, however, these algorithms are slower than the symmetric key encryption algo-

rithms and are mathematically more complex and intensive. As an example, the well-known RSA algorithm is shown in three parts as Algorithm 10, Algorithm 11, and Algorithm 12 [272–274]:

Algorithm 10 Key Generation. After [17].

```

1: procedure KEY_GENERATION( $p, q$ )
2:    $n = p * q$ 
3:   Euler's totient function,  $\Phi(n) = (p - 1) * (q - 1)$ 
4:   Choose an integer  $e$  such that it satisfies the following two conditions:
5:    $1 < e < \Phi(n)$ 
6:    $GCD(e, \Phi(n)) = 1$ 
7:   Calculate  $d$  such that  $d \equiv e^{-1}(\text{mod } \Phi(n))$ . This means  $e.d \equiv 1(\text{ mod } \Phi(n))$ 
8:    $(e, n)$  and  $(d, n)$  (public key and private key (of the receiver))
9: end procedure

```

The message to be encrypted is represented as an integer m such that $m > 0$ and m lies the interval $(0, n - 1]$. The sender has the receiver's public key (e, n) .

Algorithm 11 Encryption. After [17].

```

procedure ENCRYPT( $m, e, n$ )
  Sender computes cipher  $c = m^e(\text{ mod } n)$ .
  The cipher  $c$  is then sent to the receiver for decryption.
end procedure

```

The receiver has received the cipher from the sender.

Algorithm 12 Decryption. After [17].

```

procedure DECRYPT( $c, d, n$ )
  The cipher will be decrypted as  $m = c^d(\text{ mod } n)$ , by the receiver.
  Decrypted message is  $m$ .
end procedure

```

Time Complexity:

The complexity of the RSA algorithm is majorly contributed by three operations, which are exponentiation, inversion, and modular operation. Modular operations, such as modular addition operations exist, whose complexity is of the order of $O(\log n)$, where n is the size of the input.

Modular multiplication on 2 numbers A and B represented in k -bit binary representation, is done using squaring and multiply technique as shown in the following algorithm: To get $m^e \text{ mod } n$, modular multiplication is used. Considering the complexity of multiplication $O(\log n^2)$, i.e., repeated addition of two

Algorithm 13 Modular Multiplication using Square and Multiply Technique

```
1: procedure MOD_MULT( $A, B, k$ )
2:   Initialise output  $P = 0$ 
3:   for  $i = 0$  to  $k - 1$  do
4:      $P = 2P + A \cdot B_{k-1-i}$ 
5:      $P = P \bmod n$ 
6:   end for
7: end procedure
```

number of $\log n$ bits each, the complexity of the modular exponentiation is about $O(\log n^3)$.

Using Euclidean extended GCD from (Extended Euclidean algorithm), the inverse of a number can be calculated in $O(\log n^2)$ [275]. Thus, for N -digit number space, the overall time complexity of key generation will be $O(N^2)$, and the overall time complexity of encryption and decryption will be of the order of $O(N^3)$.

These modular operations are used repeatedly and intensively for the other cryptographic approaches also, such as Homomorphic Encryption (HE) [276]. The size of the key used should be 2048-bits and therefore, the inputs to the modular operations are also nearing the same order, which makes the overall computational overhead high.

6.7.2 Limitations of Cryptographic Approaches for Voice Privacy

Though cryptographic approaches are meant to be used for security purposes, there are practical issues with their implementations in already complex systems, such as ASVs. The limitations are discussed in this subsection.

- The security of cryptography lies under the concept of *computational* difficulty of solving the discrete logarithmic problem. However, the same reason is responsible for the limitation of cryptographic techniques in deployment to real-world applications. Therefore, cryptography is costly both in terms of time and resources.
 - Addition of cryptographic techniques in the information processing leads to delay.
 - The setting up and maintenance of cryptographic implementations, such as public key infrastructure and HE requires a large computing power, varying overhead of communications, and rounds of interactions and hence, a big monetary budget.

- Most cryptographic techniques use modulo arithmetic operations on *integers*. However, given the nature of a speech signal, representation of signals and computations on them require modulo arithmetic on *floating point* operations [114].
- Variable speech signal quality should also be reflected in the encrypted output. However, this requires computations of matrix inversions and log-determinants, which are expensive computations.
- Vulnerabilities and threats can come up because of the poor (hardware) implementation of systems, protocols, and procedures. A poor hardware implementation can open the way to many hardware-based attacks, such as side-channel attacks [277].
- Cryptographic implementations can become vulnerable to attacks if they are not maintained and updated regularly. Since the security lies in the computational difficulty, regular breakthroughs that solve those computationally difficult problems, keep coming up [17]. Hence, the current implementations should regularly update their difficulty levels.
- With the advent of quantum cryptography, the existing systems, whose security is based on the computational difficulty of solving a mathematical problem, will completely collapse. Therefore, *post-quantum* solutions in cryptography are desirable, but they too come with a cost in terms of both time and resources.

6.8 Chapter Summary

This chapter discussed privacy issues, apart from spoofing attacks, in the context of speech technologies. The design of an LP-based voice privacy system to achieve speaker anonymization is presented. Furthermore, the distribution of various speakers according to their effect on EER is discussed, which is followed by the approach of target selection from the attacker's perspective. In particular, the target selection approach is demonstrated experimentally w.r.t. twins attack. Additionally, the various technological challenges faced by the attacker are also presented. Furthermore, the merits and demerits of using standard cryptographic techniques are also discussed. In the next chapter, additional works of the proposed Morse wavelet-based features and u-vector are presented w.r.t. additional

applications on infant cry classification, and dysarthric severity-level classification.

CHAPTER 7

Additional Works

This Chapter ¹ discusses the work related to Assistive Speech Technologies (AST), such as infant cry classification and dysarthric severity-level classification. To that effect, additional use of the proposed Morse wavelet-features and the U-vector (as discussed in Chapter 5 and Chapter 4, respectively) on AST is explored in this chapter.

7.1 Infant Cry Classification

Crying is the known mode of communication for infants, which is an essential evolutionary signal that enables infants to convey discomfort. During the first three months after birth, infant cries are known to carry neurological and health status of the infant [278]. About 2.5 million infants succumb to various vaccine-preventable and other ailments within the initial months of birth. Several diagnosis methods, such as magnetic resonance imaging, computed tomography scans, and head ultrasound are used to identify various pathologies. However, pathology diagnosis is costly, and also time taking, which impacts the health of the baby. For example, asphyxia is identified by pale and bluish limbs. However, by the time these visual symptoms are visible, notable neurological damage to

¹This Chapter is based on the following publications:

- **Priyanka Gupta**, Aastha Kachhi, and Hemant A. Patil, "Classification of Normal *vs.* Pathological Infant Cries Using Morse Wavelets", submitted in 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 04 -08 Sept., 2023.
- Aastha Kachhi, Anand Therattil, **Priyanka Gupta**, and Hemant A. Patil, "Continuous Wavelet Transform for Severity-Level Classification of Dysarthria," in International Conference on Speech and Computer (SPECOM), Eds. S. R. Mahadeva Prasanna et.al, Lecture Notes in Computer Science (LNCS), Springer, vol 13721, pp. 312–324 , 2022.
- Aastha Kachhi, **Priyanka Gupta**, and Hemant A. Patil "Features Motivated From Uncertainty Principle for Classification of Normal *vs.* Pathological Infant Cry", in 30th European Signal Processing Conference (EUSIPCO), pp. 1253-1257, Belgrade, Serbia, 29 Aug. -02 Sept., 2022.

the neonate would have already occurred [279,280]. Hence, the need to develop a robust assistive pathology detection algorithm to aid paediatricians in identifying such pathologies is increasing [281]. To mitigate this problem, fingerprint-based identification systems for infants have been developed [282]. Some of the prominent causes of infant's death are asthma, Sudden Infant Death Syndrome (SIDS), and asphyxia.

The infant cry classification and analysis involves multidisciplinary fields, such as, neurological, physiological, psychology, engineering, developmental linguist, and paediatrics. Early in 1960s, the preliminary analysis of infant cries for pathological cry detection was pioneered using spectrograms by *Xie et. al.* [283]. The characteristics of normal cry sounds, generally defined as "cry modes" or "cry phonemes," including vibration, dysphonation, inhalation, and hyperphonation, were investigated for both the time and frequency-domains [283]. This study was extended to the pathological cries in [284], where some of these cry modes were found to be correlated with pathology.

In [285], infants are found to have *melodic* pattern in their cry. To that effect, a recent Constant Q Transform Cepstral Coefficients (CQCC) features were explored in [286], where the *form-invariance* property of CQT was explored for infant cry classification. Given CWT is a type of CQT, in this study, we explore the effectiveness of CWT (i.e., scalogram)-based features for infant cry classification task. Furthermore, similar to Mel-STFT, and CQT, the CWT-based scalogram features are also known to have better resolution in lower frequency regions [287,288]. Furthermore, the existing works are majorly standard neural network-based classifiers, where, the input to these neural network classifiers is fed as spectrograms, or Mel-spectrograms, which are standard and well known [289–291]. However, we have considered the standard and well-known spectrograms, Mel-spectrograms, and CQT as the baselines of our work, and Morse wavelet-based features are proposed for infant cry classification.

Furthermore, the existing works in [292] propose the use of variants of Daubechies wavelets, which are real wavelets. Moreover, these approaches use DWT-based features, whereas the analytic CWT is robust to noise and higher-order modulations, which makes it more suitable for feature extraction.

Additionally, we also exploit the u-vector feature set for infant cry classification. This work is motivated by the proposition that melody and rhythm (prosody) understanding and memory begin around the third trimester of pregnancy, and infants have a remarkable musical aptitude, where melody contours of F_0 are prominent [285]. Hence, TBP (i.e., the uncertainty in information, as discussed

in Chapter 4) can be helpful to distinguish normal *vs.* pathological cries based on the information from the F_0 contours. To that effect, we propose uncertainty-based feature set called as *u-vector* for the classification of normal *vs.* pathological cries.

7.1.1 Morse wavelets-Based Features

As per the study reported in [283], an infant cry signal comprises 10 cry modes. These cry modes have different time-frequency patterns, when observed in narrowband spectrogram. These cry modes (as shown in Figure 7.1 (a)) are:

1. **Glottal Roll:** Also called as the *trailing* cry mode, it comprises gradually decreasing patterns of F_0 and total energy.
2. **Flat:** It is the region where F_0 is smooth and steady with fewer variations between F_0 and its harmonics are observed.
3. **Falling:** It is the time-frequency region having descending F_0 .
4. **Rising:** It is the time-frequency region having ascending F_0 .
5. **Double Harmonic Break:** It represents the weak primary simultaneous parallel harmonic lines present between harmonics of F_0 .
6. **Dysphonation:** It represents the regions of the time-frequency representation, where the harmonics are indistinguishable, and the energy distribution is *unstructured* over all the frequencies.
7. **Hyperphonation:** Energy distribution with high F_0 phonation.
8. **Inhalation:** It is caused by rapid breathing of the infant, resulting in exhaustive expiratory phase.
9. **Vibration:** Normally time-frequency pattern of high energy level with unstructured energy distribution of vibrating F_0 .
10. **Weak Vibration:** It is similar to vibration, but with lower energy levels of F_0 .

It should be noted that continuous-time Fourier transform (CTFT) obeys the form-invariance property, i.e., $\mathcal{F}\{x(t)\} = \frac{1}{|\alpha|} X\left(\frac{\omega}{\alpha}\right)$. However, STFT does not obey form invariance (because the window function of the time parameter), as obeyed by CQT, and CWT. To that effect, we can observe in Figure 7.1 (b) and Figure 7.1

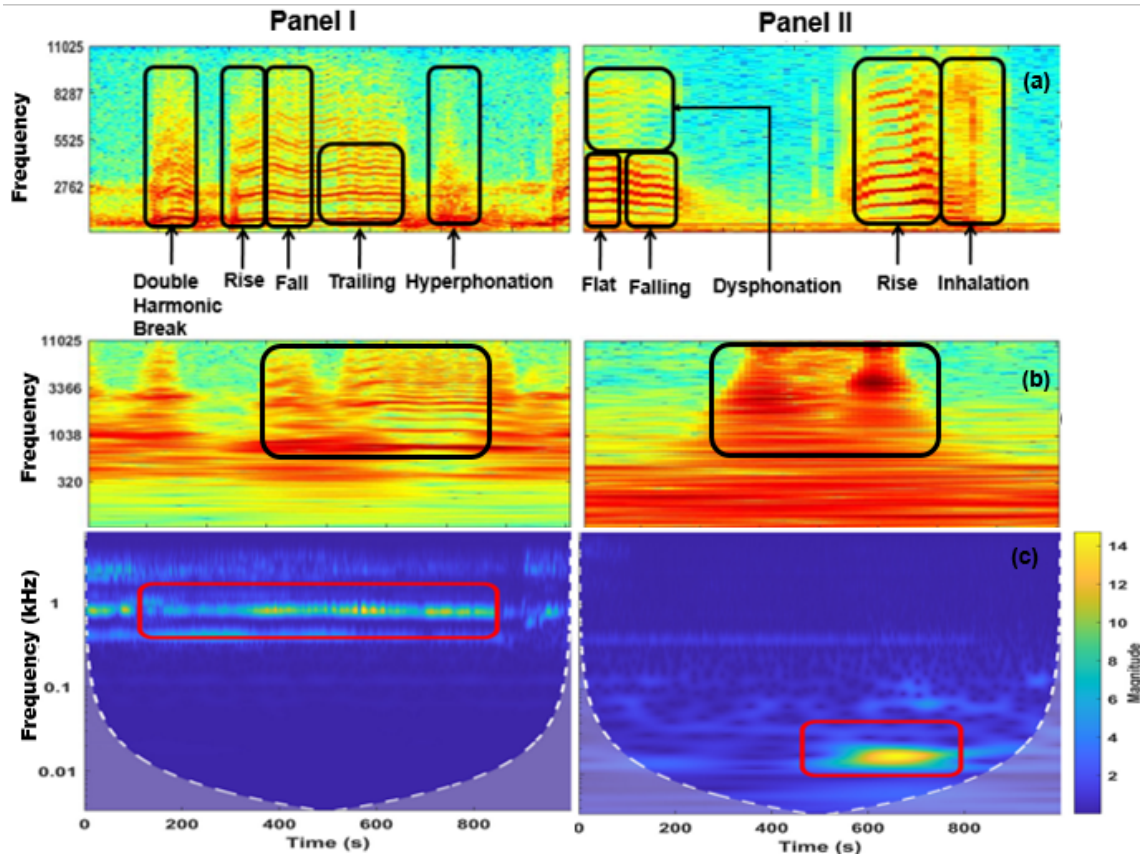


Figure 7.1: Infant cry modes captured by (a) spectrogram, but have a general structure in (b) CQT-gram, and (c) Morse wavelet-based scalogram representations, due to the form-invariance property followed by CQT and CWT.

(c) that these cry modes have a common (i.e., general) structure in CQT-gram and scalogram representations, respectively. Since CWT is also a kind of CQT, the scalogram also follows the form-invariance property and hence, a common structure is observed.

With respect to the discussion in Chapter 5, similar Morse wavelet-based scalograms are used as features for infant cry classification in this work. However, in Chapter 5 for VLD, we used low-frequency region (0-50 Hz) for feature extraction, whereas in this chapter, we have used the entire frequency range from 0 to $f_s/2$ Hz. From Panel-I of Figure 7.1 (c), it can be observed that the scalogram corresponding to healthy (normal) cry exhibits a *continuous* band of energy in the high frequency region (particularly near 1 kHz, as shown by the red box). However, a clear distinction can be observed in Panel-II of Figure 7.1 (c), which shows scalogram corresponding to pathological (asphyxia) infant cry. In particular, it can be observed that there is *absence* of continuous band of energy in this case. Moreover, a *short burst of energy* (as indicated by the red box) is observed in the lower frequency region, indicating the possible presence of pathology. In this work, we

Algorithm 14 MATLAB pseudo code for the proposed Morse wavelet-based scalogram extraction.

```

1: procedure SCALO_IMAGE( $x$ )                                ▷  $x$  is the speech signal
2:    $w\_name = 'morse'$                                        ▷ Taking Morse wavelet
3:    $[cwt\_coeffs, F] \leftarrow cwt(x, w\_name)$ 
4:   Mapping CWT coefficients to scalogram RGB image of  $512 \times 512$ .
5:    $tpoints \leftarrow$  time vector of  $x(t)$ 
6:    $pcolor(tpoints, cwt\_coeffs)$ 
7:    $scimg = ind2rgb(rescale(cwt\_coeffs), jet(320))$ 
8:    $scalo\_image \leftarrow imresize(scimg, 512, 512)$ 
9: end procedure

```

extract CWT-based representations (i.e., scalograms), as described in Algorithm 14. These scalogram-based features are then fed as input to the GMM classifier.

7.1.1.1 Experimental Setup

- **Datasets Used:** In this study, we use three datasets, namely, Baby Chillanto database, the DA-IICT corpus, and the combined (Baby Chillanto + DA-IICT) dataset. Baby Chillanto dataset was designed using recordings made by doctors and is the property of Mexico’s NIAOE-CONACYT and its statistics can be studied from [293]. The second dataset known as the DA-IICT corpus was collected by [294, 295]. The third dataset called as combined dataset consists of all utterances of Baby Chillanto dataset along with DA-IICT corpus, with 1842 utterances in the healthy class, and 1616 utterances in the pathology class. For fair experimentation, we have resampled the utterances of all the datasets to a uniform sampling frequency of 16 kHz. The resampling was done using MATLAB’s resample function, which resamples the input signal by applying an FIR Antialiasing Lowpass Filter to the signal and compensates for the delay introduced by the filter. The function operates along the first array dimension with a size greater than 1. Experiments in this study are performed using 10-fold cross-validation (cv) method. Table 7.1 shows the data partitioning of each of the 10-folds.

7.1.1.2 Experimental Results

- **Effect of wavelet duration ($P_{\beta,\gamma}^2$):** Here, we present the results obtained by varying the wavelet duration parameter of the Morse wavelet, $P_{\beta,\gamma}^2$. Figure 7.2 shows the effect of $P_{\beta,\gamma}^2$, keeping γ fixed as 3, on the two

Table 7.1: Number of utterances in data partitions in each fold.

Dataset →	Baby Chillanto	DA-IICT	Combined
Train	2041	1071	3112
Test	227	119	346
Total	2268	1190	3458

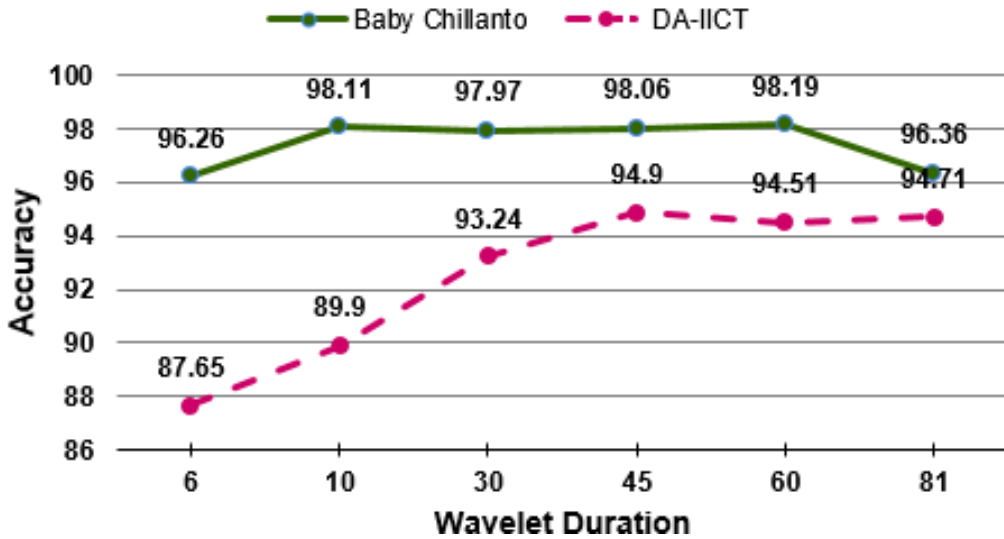


Figure 7.2: Effect of the Morse wavelet parameter $P_{\beta,\gamma}^2$ on Baby Chillanto, and DA-IICT corpora.

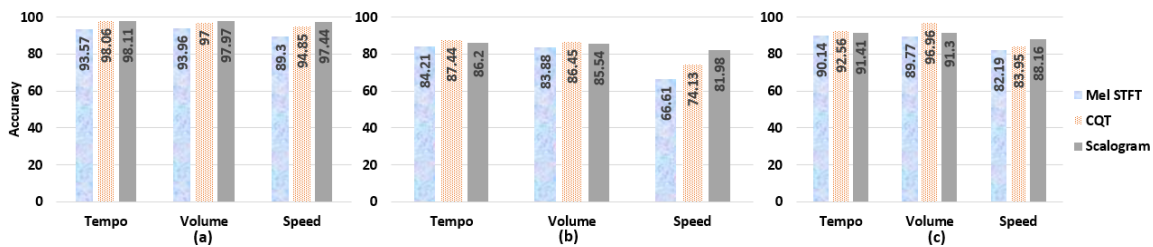


Figure 7.3: Effect of three data augmentation techniques (i.e., tempo, volume, and speed perturbations), on (a) Baby Chillanto, (b) DA-IICT, and (c) combined corpora. Best viewed in color.

datasets. It can be observed from Figure 7.2 that the results are degraded for relatively lower values of wavelet duration, more so, for the DA-IICT corpus. From the discussion in Chapter 5, it can be deduced that for a fixed value of γ (i.e., $\gamma = 3$ in this work), a low value of $P_{\beta,\gamma}^2$ corresponds to a low value of β . Therefore, in Figure 5.17, as we keep $\gamma = 3$ fixed, and move towards lower values of β (i.e., lower values of wavelet duration, $P_{\beta,\gamma}^2$), poor quality filters (with low quality factor, i.e.,

with increased bandwidth) are produced. Therefore, the performance is observed to be degraded for lower values of wavelet duration for both the datasets. Furthermore, the optimal value of $P_{\beta,\gamma}^2$ obtained is 60 for the case of Baby Chillanto dataset, and 45 for the case of DA-IICT corpus. Given that Baby Chillanto is statistically significant standard dataset for infant cry classification problem, we choose to keep $P_{\beta,\gamma}^2$ as 60, for the remaining experiments in this work.

- **Overall Performances:** Keeping the parameter $P_{\beta,\gamma}^2$ as 60 for Morse wavelet-based scalogram features, we now present the overall performances in terms of % accuracy, obtained on the three datasets used in this work. To that effect, Table 7.2 shows the performance of Mel STFT, CQT, and the proposed Morse wavelet-based scalogram features. It

Table 7.2: Overall performance (in % accuracy) of baselines and the proposed features on the three datasets.

Datasets →	Baby Chillanto	DA-IICT	Combined
Mel-STFT	92.16	87.45	89.57
CQT	97.31	89.41	92.05
Morse wavelet -based scalogram	98.19	94.51	91.70

can be observed that the proposed features outperform Mel STFT and CQT features on Baby Chillanto and DA-IICT corpora. On the Baby Chillanto dataset, the absolute improvement in % accuracy is 6.03 and 0.88, w.r.t. Mel STFT and CQT, respectively. On the DA-IICT dataset, the absolute improvement in % accuracy is 7.06 and 5.1, w.r.t. Mel STFT and CQT, respectively. Furthermore, for the combined dataset, CQT still remains to be performing the best. However, the absolute difference of % accuracy of CQT and Morse wavelet-based scalogram features is only 0.35. The marginally improved performance of CQT might be due to the data partitioning effect on the combined data.

- **Effect of Data Augmentation:** Here, the results (in % accuracy) pertaining to the three data augmentation techniques are presented, namely, tempo, volume, and speed perturbation. To that effect, Figure 7.3 (a) shows the results pertaining to Baby Chillanto dataset, wherein the performance on three types of data augmentation techniques (tempo, volume, and speed perturbation) is shown. It can be observed that Morse wavelet-based scalogram approach outperforms the existing ap-

proaches for *all* the three augmentation techniques in the case of Baby Chillanto dataset. Furthermore, Figure 7.3 (b) shows the results pertaining to DA-IICT dataset. It can be observed that CQT-based method is observed to perform the best for tempo and volume perturbation (with accuracies of 87.44% and 86.45%, respectively), whereas for speed perturbation, Morse wavelet-based features show the best performance of 81.98% accuracy. A similar trend can be observed in Figure 7.3 (c), which shows the results pertaining to the combined dataset.

It is worth noting that *all* the datasets, the performance of speed augmentation is relatively lower than tempo and volume perturbation for each of the Mel STFT and CQT-based features. However, the performance degradation is negligible for the proposed Morse wavelet-based scalogram feature set.

7.1.2 Uncertainty Feature Vector

The uncertainty feature vector is based on the Heisenberg's uncertainty principle, to capture the uncertainty in the cry signal in signal processing framework, with details discussed in Chapter 4. The u -vector is constructed with the help of two other feature vectors, namely, t -vector and ω -vector, which are discussed in detail in Chapter 4. These vectors represent the variance in time and frequency-domains, respectively. The infant cries contain music-like harmonics, and in order to evaluate the domain (time or frequency) in which the variability plays a more contributing role, the variance in time and in frequency domain are measured, which leads to the formation of the uncertainty vector. Further, the latency period analysis for u -vector, t -vector, and ω -vector features is also shown in this work. The significance of this study is further strengthened by information-theoretic measures, namely, KLD and JSD (as discussed in subsection 4.2.5.11 of Chapter 4).

The proposed feature extraction for infant cry classification is based on the fact that the spectral energy density patterns are different for healthy *vs.* pathological cries. This is also shown by the spectrographic analysis in Figure 7.4, which shows that pathological cries have high frequency of inhalation, indicating problems while breathing. Hence, spectral smearing is found in the entire frequency range. It can also be observed that there is a sudden rise in the pitch (F_0) source harmonics and smearing in some regions. Therefore, the frequency variance helps to capture the regions of

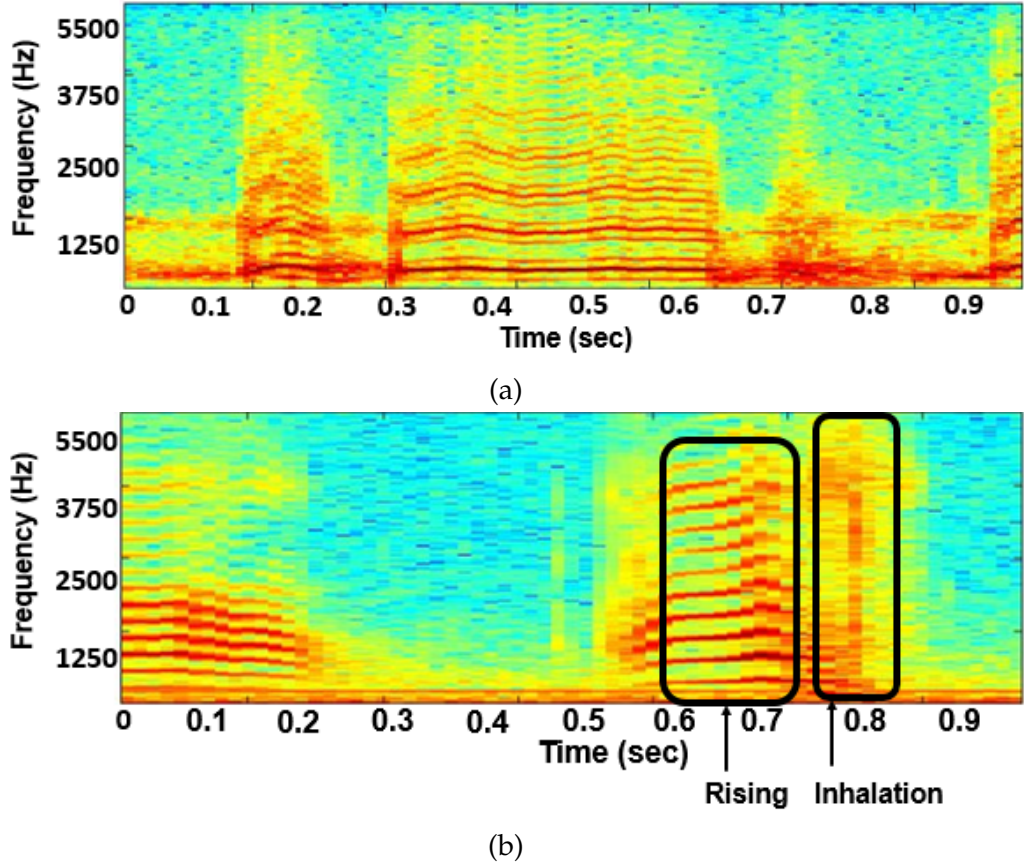


Figure 7.4: Spectrograms of (a) healthy *vs.* (b) pathological cries.

spectral smearing.

The feature extraction procedure in this work begins by passing the cry signal $s(t)$ through a Gabor filterbank of 40 subband filters. This results in 40 subband signals $s_i(t)$, where $i \in [1,40]$. Since the infant cry signal is multi-component, the subband signals help in capturing frequency variances effectively [217]. Here, 40 linearly-spaced Gabor filterbank is used because of its *optimal* time and frequency resolution [10,84]. Each of the subband output signals is frame-blocked with a window size of 30 ms and window shift duration of 15 ms (which is experimentally optimized). For each of these frames, both σ_t^2 and σ_ω^2 is computed and hence, three different vector representation of the input speech signal are obtained as shown in Algorithm 15. Next, logarithmic operation is then performed on σ_t^2 and σ_ω^2 to extract t -vector and ω -vector of the infant cry signal. Similarly, logarithm on the product $\sigma_t^2 \sigma_\omega^2$ gives the u -vector or the uncertainty vector of the cry signal, as indicated by eq. (4.53) and eq. (4.54).

Algorithm 15 TBP Computation for Infant cry.

```
1: procedure U-VECTOR( $x$ ) ▷  $x$  is the speech signal
2:    $T \leftarrow$  Gabor filterbank ( $x$ )
3:   Window length = 30 ms, window overlap = 15 ms
4:   for  $j=1$ :number of frames do
5:      $Var_t \leftarrow$  variance( $T(j, :)$ , mean) ▷  $t$ -vector
6:      $mean_f \leftarrow$  mean(FFT( $T(j, :)$ ), freq)/(2 *  $\pi$ )
7:      $Var_f \leftarrow$  variance( $A$ ,  $mean_f$ , freq) ▷  $\omega$ -vector
8:      $tbp_{gen} \leftarrow var_t * var_f$  ▷  $u$ -vector
9:   end for
10:  return  $tbp_{gen}$ 
11: end procedure
```

7.1.2.1 Experimental Setup

- **Dataset Used:** We use Baby Chillanto database, with the statistics as shown in Table 7.1. Each recording was segmented to make infant cry signals of 1 second duration each. Since the sampling rate of the cry signals provided in the dataset is not uniform, we resampled all the utterances at a sampling rate of 11.025 kHz.
- **Classifier used:** The experiments were performed using Gaussian Mixture Model (GMM) classifier, which is commonly used for infant cry classification [189,296]. In this study, 512 mixture components are used to train the model.
- **Baseline:** We consider CQT feature as the baseline for this work [297], where for low frequency regions, CQT gives better frequency resolution. For this baseline, we performed experiments with 96 number of bins per octave, keeping $f_{min} = 100$ Hz.

7.1.2.2 Experimental Results

This work is performed using *10-fold* cross-validation on Baby Chillanto dataset. We performed the experiments by fine-tuning feature parameters, such as window overlap and number of subband filters.

- **Effect of window overlap:** We varied the window overlap with values as 15, and 20 ms. The number of subband filters was kept constant. The obtained experimental results are presented in Table 7.3, which shows that the highest performance is achieved as 93.83% accuracy, obtained

Table 7.3: % Accuracy for Non-Cepstral and Cepstral u -vector.

Window Length	Window Overlap	# Filters	% Accuracy (non-cepstral)	% Accuracy (cepstral)
30	15	40	93.83	93.04
30	15	60	93.08	93.48
30	15	80	87.71	87.00
30	20	40	93.35	92.42

when window length, window overlap, and number of subband filters are of 30 ms, 15 ms, and 40, respectively.

- **Effect of number of subband filters:** Further, the next set of experiments was performed by varying the number of subband filters and keeping window overlap constant. These fine-tunings were performed considering the two cases of non-cepstral, and cepstral u -vector. It should be noted that the non-cepstral u -vector (with 93.83% accuracy) performs better than its cepstral version (with 93.48% accuracy). It can also be observed from Table 7.3 that as the number of subband filters increases, the % classification accuracy decreases.
- **Comparison of u -vector, t -vector, and ω -vector with the CQT:** Table 7.4 shows the comparison of the performances of u -vector, t -vector, and ω -vector with the CQT baseline. The comparison is done for both the cases of cepstral and non-cepstral features. It can be observed that the

Table 7.4: % Classification Accuracy for Various Cepstral and Non-Cepstral Feature Set.

Non-Cepstral Features		Cepstral Features	
Feature Set	% Accuracy	Feature Set	% Accuracy
u -vector	93.83	u -vector	93.48
t -vector	91.23	t -vector	89.38
ω -vector	98.50	ω -vector	96.74
CQT	97.00	CQT	98.55
Average	95.14	Average	94.53

non-cepstral ω -vector performs the best with % classification accuracy of 98.5% with overall increase of about 1.5% than the baseline CQT feature. Hence, it can be observed that the frequency distribution patterns of the different cry modes smeared over the entire frequency band, are captured by the ω -vector as discussed in [283]. Further, it can be observed that out of all the features shown in Table 7.4, the best perfor-

mance is achieved by ω -vector in the non-cepstral case with an accuracy of 98.50%. Furthermore, it should also be noted that the average overall accuracy of non-cepstral feature is higher than the cepstral features. In particular, the non-cepstral features achieve average higher accuracy (95.14%) as compared to the cepstral features. This indicates that non-cepms after the burststral features are better suited for pathology detection.

Given that ω -vector achieves the best performance in the non-cepstral domain, we performed the next set of experiments to observe the effect of number of subband filters in the ω -vector. Table 7.5 presents the

Table 7.5: % Classification Accuracy of ω -vector with Various Number of Subband Filters.

Subband Filters	30	40	60	80	100
% Accuracy	92.03	98.50	91.37	92.20	96.78

corresponding results, and it can be observed that the best result 98.50% is achieved with 40 number of subband filters. From Table 7.5, we can say that when the entire frequency band is divided into 40 subbands, the frequency variance captured in each subband is optimum for our binary classification task.

- **Model-level measure of discriminative ability:** To estimate the model-level measure of discriminative ability w.r.t. various feature sets, we use KLD and JSD. Figure 7.5 shows KLD and JSD of non-cepstral and cepstral features in Panel-I and Panel-II, respectively. It can be observed that for the case of non-cepstral features as shown in Panel-I, ω -vector outperforms all the remaining features sets in terms of KLD as well as JSD. This shows that the better discriminative ability of our model is achieved when ω -vector is used. This is also reflected in the % accuracy results achieved, shown in Table 7.4, where ω -vector outperforms all the remaining features in the case of non-cepstral features. Furthermore, for the case of cepstral features as shown in Panel-II, it can be observed that CQT shows better discriminative ability for the case of KLD (healthy || pathology), and JSD (healthy || pathology). However, it should be noted that ω -vector shows better performance for the case of KLD (pathology || healthy).
- **Analysis of Latency Period:** We also investigate the latency period for t -vector, ω -vector, and u -vector w.r.t the CQT feature set. The latency is

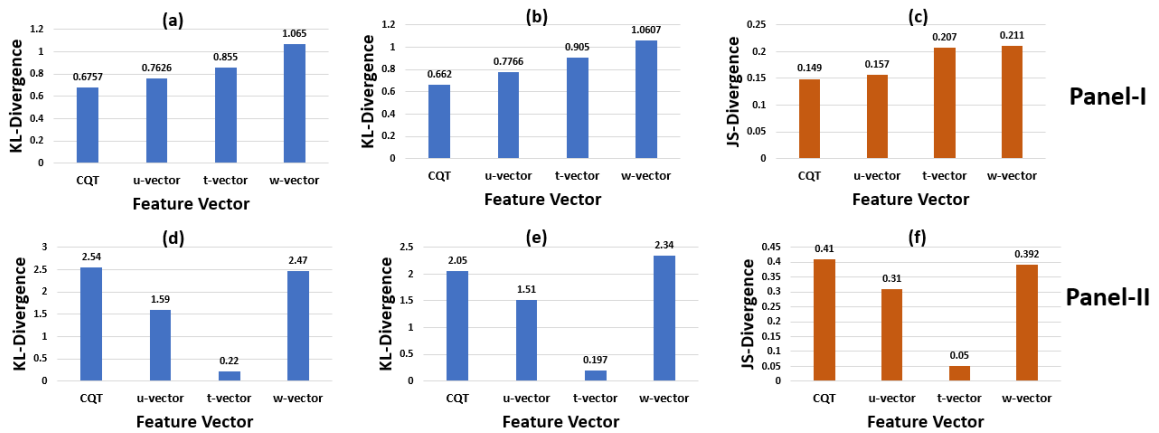


Figure 7.5: KLD and JSD of the proposed feature sets. Panel-I and Panel-II denote the cases of non-cepstral features and cepstral features, respectively. KLD (healthy || pathology) is shown in (a) and (d). KLD (pathology || healthy) is shown (b) and (e). JSD (healthy || pathology) is presented in (c) and (f).

estimated by the performance evaluation in terms of % accuracy w.r.t. varying durations of speech segment in an utterance. The duration of the utterance ranges from 20 ms to 600 ms, with an interval of 150 ms. We have estimated KLD and JSD on GMMs of 512 mixtures, for each of the feature vectors. Figure 7.6 shows comparison between non-cepstral features of CQT, u -vector, t -vector, and w -vector. It can be observed that the w -vector outperforms u -vector and t -vector, and shows remarkable latency as compared to the CQT. Moreover, it can be observed that all three features, i.e., u -vector, t -vector, and w -vector gave increased % accuracy in a short duration of speech utterance of < 200 ms. On the other hand, CQT showed no improvement in accuracy even for a long duration of 600 ms of a speech utterance. Additionally, the feature performance is better if for a low latency period the accuracy is high, which indicates the faster classification by the model and thus, indicates suitability for practical infant cry classification system deployment.

7.2 Dysarthric Severity-Level Classification

Proper coordination between the brain and speech-producing muscles is required for the production of speech sounds [298]. Lack of this coordination leads to speech disorders, such as apraxia, dysarthria, and stuttering. These disorders affect a person's ability to produce speech sounds. They are further categorized as neurological or neurodegenerative diseases, such as

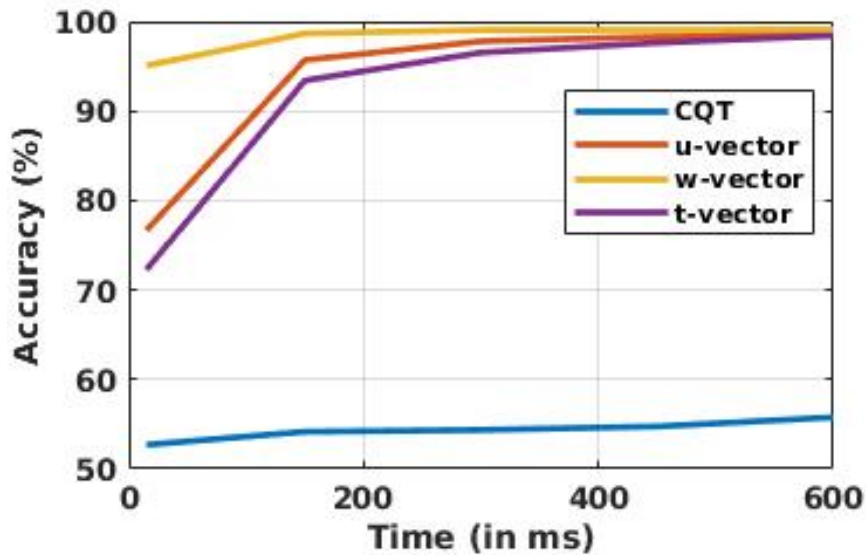


Figure 7.6: Latency period *vs.* % accuracy between the various non-cepstral features for CQT, u -vector, t -vector, and ω -vector.

cerebral palsy or Parkinson’s disease. The severity level of these diseases might be mild or severe, depending upon the impact on the area of the brain. In the case of mild severity, the patient may mispronounce a few words, whereas, in high severity, the patient lacks the ability to produce intelligible speech. Among these speech disorders, dysarthria is a relatively common speech disorder [299]. Dysarthria is a neuromotor speech disorder. The muscles that produce speech are weak in people with this disorder. Dynamic movements of articulators, such as lips, tongue, throat, and upper respiratory tract system are also affected due to brain damage. Apart from brain damage, cerebral palsy, muscular dystrophy, and stroke are also some of the other factors, which can cause dysarthria [300].

Severity-level of dysarthria depends on the impact and damage to the area of neurological injury, which is diagnosed using a brain and nerve test. The type, underlying cause, severity-level, and its symptoms, all influence the manner in which it is treated [301]. Due to this uncertainty in treatment, researchers are motivated to develop speech assistive tools for dysarthric intelligibility categorization.

In the literature, dysarthria severity-level classification has been exploited extensively using STFT [302], and various acoustical features [303]. Standard feature sets, such as MFCC were employed in [304] due to their capacity of capturing global spectral envelope properties. In addition to a perceptually-motivated state-of-the-art feature set, glottal excitation source

parameters derived from the quasi-periodic sampling of the vocal tract system were implemented in [305]. In the signal processing framework, due to the wide and dynamic range of multiple frequency components in short-time spectra, speech signals are considered to be non-stationary signals. Due to the dynamic movements of articulators, the frequency spectrum varies instantaneously.

7.2.1 Morse wavelet-based features

In this work, we demonstrate the capability of CWT-based representation (i.e., scalogram) for dysarthric severity-level classification. The key motivation of utilizing CWT for this study is the improved frequency resolution of CWT-based scalograms at lower frequencies as compared to the STFT-based and Mel spectrogram-based techniques. To the best of the author’s knowledge and belief, the use of CWT has been explored to Model Articulation Impairments in Patients with Parkinson’s Disease [306]. However, the use of CWT to capture discriminative acoustic cues for dysarthric severity-level classification is being proposed for the first time in this thesis work. Results are presented on standard Universal Access (UA)-Speech Corpus. In this work, scalogram images were extracted using MATLAB with $\gamma=3$ and $\beta=20$ (i.e., $P_{\beta,\gamma}^2=60$) as the default parameter setting for Morse wavelet-based scalogram for full frequency band up to 8 kHz (since sampling frequency, $F_s = 16$ kHz). Each scalogram image extracted is of $512 \times 512 \times 3$ dimension. These scalogram-based features are then fed as input to the CNN classifier. The experimental setup is explained in the following subsection.

Panel I of the Figure 7.7 shows the speech segment of vowel /e/. Panel II, III, and IV show the spectrogram, Mel spectrogram, and scalogram, respectively, for (a) normal, (b) very low, (c) low, (d) medium, and (e) high dysarthric severity-level for the same speech segment. It can be observed from Figure 7.7 that the scalogram-based features can capture energy-based discriminative acoustic cues for dysarthric severity-levels more accurately than the STFT and Mel spectrogram-based features. Furthermore, from scalogram, it can be observed that as the dysarthtic severity-level increases, patients struggle to speak the prolonged vowel, /e/. This may be due to the lack of coordination between articulators and the brain. Due to this, the energy spread is seen over the entire time-axis. However, the utterance of vowel /e/ is of short duration for medium and high dysarthtic severity-

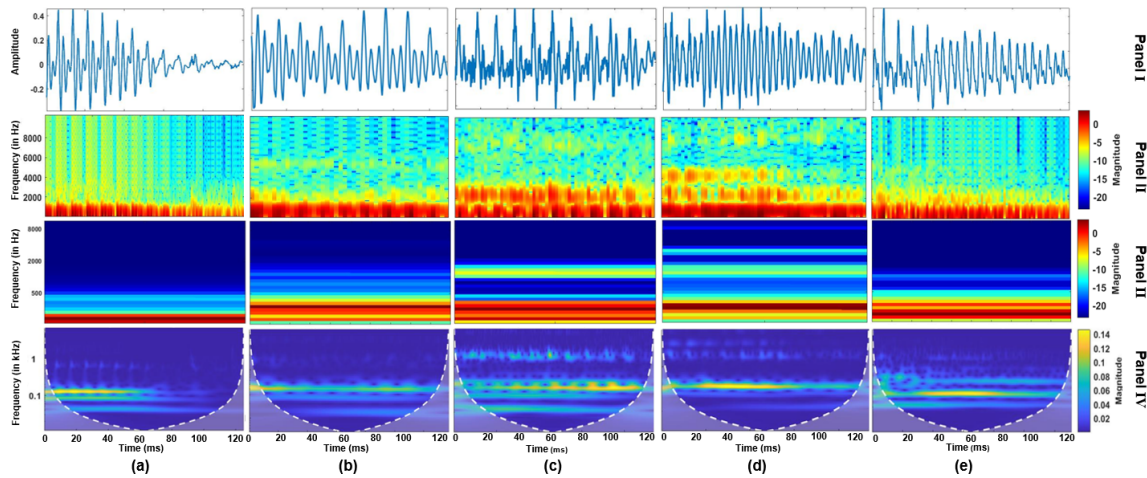


Figure 7.7: Dysarthric speech utterance (for vowel /e/) for male speaker with various dysarthric severity-level (Panel I), corresponding STFT (Panel II), corresponding Mel spectrogram (Panel III), and corresponding Morse wavelet scalogram (Panel IV) for (a) normal, dysarthric speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. Best viewed in color.

levels.

7.2.1.1 Experimental Setup

- **Dataset Used:** The Universal Access dysarthric Speech (UA-Speech) corpus [307] is used to evaluate the proposed CWT-based approach. In this study, a dataset configuration identical to that described in [302] is used. It has 8 speakers, out of which 4 are male and 4 are female speakers. Furthermore, 90% of the dataset is dedicated to training set and the remaining 10% is dedicated to the testing partition.
- **Classifier Used:** Based on the experiments presented in [304], the Convolutional Neural Network (CNN) is used as a classifier in this study. According to a study reported in [304], CNN gives comparable results with the other deep neural network (DNN)-based classifiers for the UA-Speech corpus. For this study, the CNN model was trained employing the Adam optimizer algorithm, four convolutional layers with kernel size of 5×5 , and one Fully-Connected (FC) layer [156]. Mel spectrograms and scalograms, both of size 512×512 , were used in these investigations. A max-pool layer and Rectified Linear Activation (ReLU) are utilised. For loss estimation, a learning rate of 0.001 and cross-entropy loss are chosen.

7.2.1.2 Experimental Results

The performance evaluation for various feature sets is done *via* % classification accuracy (as shown in Table 7.6). On CNN, the scalogram performs relatively better with a classification accuracy of 95.17% than the baseline STFT, and Mel spectrogram. Furthermore, Table 7.7 shows the confusion

Table 7.6: Results in (% Classification Accuracy) for CNN Classifier.

Feature Set	CNN
STFT	91.76
Mel-Spectrogram	92.65
Scalogram	95.17

matrix of the STFT, Mel spectrogram, and Morse wavelet-based scalogram for CNN model. It can be observed that the scalogram reduces the false prediction error, which indicates the better performance of the scalogram *w.r.t* the baseline STFT, and Mel spectrogram.

Table 7.7: Confusion Matrix Obtained for STFT, Mel-Spectrogram, and Scalogram.

Feature Set	Severity	High	Medium	Low	Very Low
STFT	High	63	6	3	3
	Medium	10	79	3	1
	Low	3	4	79	7
	Very Low	1	2	1	89
Mel-Spectrogram	High	69	1	3	2
	Medium	5	81	4	3
	Low	4	1	91	0
	Very Low	4	0	2	89
Scalogram (Morse Wavelet)	High	69	5	1	0
	Medium	3	89	1	0
	Low	1	1	90	1
	Very Low	3	0	1	89

The capabilities of scalogram for the classification of the dysarthric severity-level is also validated by LDA scatter plots due to its higher image resolution and better projection of the given higher-dimensional feature space to lower-dimensional space than the scatter plots obtained using t-SNE plots [308]. Here, the LDA plot of STFT, Mel spectrogram, and scalogram are projected onto 2-D feature space, and represented using the scatter plot shown in Figure 7.8 (a), Figure 7.8 (b), and Figure 7.8 (c), respectively. From Figure 7.8,

it can be observed that wavelet-based scalogram has low intra-class variance and high inter-class variance, which increases the distance between the clusters *w.r.t* baseline STFT, and Mel spectrogram, thereby better classification performance by the proposed Morse wavelet-based approach.

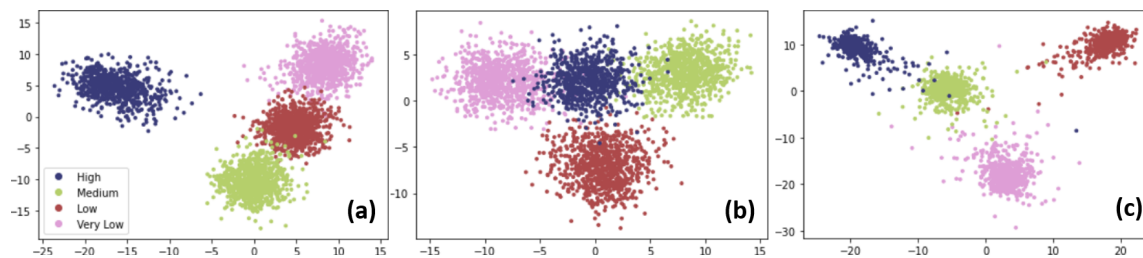


Figure 7.8: Scatter plot obtained using LDA for (a) STFT, (b) Mel spectrogram, and (c) Scalogram. Best viewed in color. After [18].

7.3 Chapter Summary

This Chapter explored additional applications of two of the proposed features sets, namely, Morse wavelet-based feature set, and the u-vector. To that effect, both these feature sets were explored for the problem of infant cry classification. For the case of Morse wavelet-based features, the performance was evaluated on three datasets, namely, Baby Chillanto, DA-IICT, and combined dataset. Additionally, the effect of three data augmentation techniques (tempo, volume, and speed perturbation) was also discussed. For the case of u-vector, it was observed that non-cepstral features are better suited for detection of pathological cries. Given the early detection of pathology in infants is also associated with faster detection, the latency period performance of the proposed features was also analyzed. Furthermore, another application of dysarthric severity-level classification was explored *w.r.t*. Morse wavelet-based scalogram features. It was observed that the energy spread corresponding to the dysarthric severity in low-frequency region is better visualized in the scalogram. Hence, the low-frequency discriminative cues are better classified using a scalogram. In the next chapter, the summary of thesis is discussed in brief, along with the limitations and future scope, and open research problems derived from this thesis work.

CHAPTER 8

Summary and Conclusions

This chapter summarises the work presented in this thesis and discusses its limitations, prospective future study areas, and a few unresolved (or open) research problems.

8.1 Summary of the Thesis

This thesis began with an introduction to ASV systems and the need to design CM solutions to counteract potential spoofing attacks on ASV systems in Chapter 1. Next, Chapter 2 presented the literature survey on replay spoof detection, VLD, and the attacker’s perspective. Furthermore, Chapter 3 presented the experimental setup used for various experiments that are reported in this thesis. The setup includes the details of the various datasets, classifiers, and performance metrics used. Chapters 4, 5, and 6 include the major contributions of this thesis towards defense against spoofing attacks. Following these, Chapter 7 showed two additional applications of the proposed features on infant cry classification, and dysarthric severity-level classification.

Chapter 4 discussed the proposed handcrafted features for the replay SSD task. To that effect, the CFCCIF-QESA feature set was predominantly discussed in this chapter, followed by two additional features, namely, optimized LFRCC and u-vector. The CFCCIF-QESA feature set is an improvement to the recently proposed the CFCCIF-ESA feature set, which additionally incorporates the quadrature-phase component. The incorporation of the quadrature phase enables capturing additional information in the signal, which further improves the performance of the SSD system. The CFCCIF-ESA feature set utilizes only the amplitude information of the three

consecutive speech samples. Moreover, due to the absence of the Hilbert transform, it does not utilize the quadrature-phase component of the signal for analytic signal generation. Therefore, in order to incorporate both the advantages, CFCCIF-QESA was developed in this thesis. The quadrature-phase component is included by proposing QESA, which uses the existing ESA with an extended definition of TEO for complex signals. So far, TEO for real-valued signals has been used extensively in the literature for replay SSD task. This thesis proposes the use of the extended definition of TEO for complex-valued signals for the replay SSD task. Furthermore, a comprehensive study of the various methods used to estimate IF is followed by a discussion of the significance and justification of the selection of the quadrature-phase component as well as the in-phase component by MI-based analysis, which provides the justification of incorporating the quadrature-phase component for the design of the CFCCIF-QESA feature set. Additionally, the proposed QESA method is used to offer a thorough discussion on the IF difficulties w.r.t. to the elimination of a few of the difficulties.

The discussion on CFCCIF-QESA is followed by a discussion and results on the optimized LFRCC feature set. The LP residual is known to capture discriminating information for the replay SSD task. Hence, the effect of LP order on the residual is analyzed. Notably, the information carried by the LP residual also depends on the LP order, p . A relatively large value of the order will lead to good prediction of speech signal and hence, lower error (i.e., LP residual), and vice-versa. However, for SSD task, our aim is not to have a good prediction of speech, rather to exploit the residual at an order optimally suited for the SSD task. Therefore, LP order for the replay SSD task has been found experimentally as 8 using the ASVSpooF 2019 PA dataset. In particular, lower LP order means poor prediction of speech and thus, LP residual will sound more intelligible, indicating the characteristics of genuine speech, and thus, it will help to discriminate more vividly the characteristics of spoofed speech. This means that for LP order 18, we have relatively the best possible prediction of speech, and thus, its spectrum (in particular, formant peaks), more clearly, because the roots of LP predictor are used for speaker anonymization. Therefore, LP order used for the design of voice privacy system in Chapter 6 was taken to be 18 for speech utterances with $f_s = 16$ kHz. Another feature vector, namely, u-vector has been proposed for replay SSD. It is based on capturing the area of the Heisenberg's box, which is known to capture the richness of information in a signal. The

u-vector comprises two other feature sets, which are also evaluated for the replay SSD task, namely, t-vector and ω -vector. The time variance (σ_t^2) and frequency variance (σ_ω^2) of the signal are used for the extraction of t-vector and ω -vector, respectively.

The reliability of the existing SSD systems on a specific attack type prevents them from being designed as a generalized SSD system when taking into account the real-world scenario, where an attacker is an external entity free to choose any technique of generating the spoofed signal. This is majorly due to the fact that the existing SSD systems rely on the characteristics of the spoofed signal to detect whether a speech utterance is genuine or spoofed. Therefore, VLD systems are a step towards alleviating this issue, by using the characteristics of live speech instead of spoofed speech. To that effect, current VLD systems exploit pop noise as a discriminative acoustic cue to detect whether the speech is live or not. The VLD task is based on pop noise detection, which is an acoustic cue produced by the sudden burst of air on the microphone caused by the proximity of the speaker's mouth and the microphone, and is present in very low-frequency regions. To that effect, the high-frequency resolution of the CWT in the lower-frequency regions enables us to capture the pop noise cues effectively. In this context, Chapter 5 presented three analytic wavelets-based features for the VLD task, namely, Bump wavelet-based, Morlet wavelet-based, and GMW-based feature sets. The experimental results were presented in the Chapter along with distance-based analyses. In particular, GMWs are shown to be a superfamily of analytic wavelets, and hence, much detailed experiments and analysis are shown w.r.t. Morse wavelet-based features for the VLD task. In particular, while the experiments on bump and Morlet wavelet-based features focused only on the performance of the VLD system w.r.t. the existing STFT-based approach, the experiments on Morse wavelet-based features additionally include detailed experiments, such as the effect of wavelet parameters, attacker-speaker distance, and speaker-microphone distance.

Chapter 6 discusses some of the attacker's perspectives w.r.t. voice privacy. Given a speech signal carries paralinguistic information, such as gender, age, health, and emotional status, ethnicity of a speaker, the speaker's identity can be under threat. To that effect, voice privacy aims at hiding a speaker's identity while keeping linguistic content and naturalness of the speech intact. To that effect, a modification to the *baseline-2* of the Voice Privacy Challenge 2020 is presented in Chapter 6, which uses LP analysis to

hide speaker's identity. Furthermore, attacker's perspective w.r.t. choosing the most vulnerable speaker using the approach of target selection is presented. To that effect, target selection on twins is performed to find out the most vulnerable twin-pair, i.e., the twin-pair which has the highest chances of succeeding in a twins attack. The relevance of VP system in protecting a speech corpus from target selection is also discussed.

8.2 Limitations of This Work

Limitations of this thesis work are as follows:

- One of the limitations of the proposed IF estimation in CFCCIF-QESA is that it does not alleviate all the difficulties associated with IF definition and estimation and hence, this remains an open research question.
- The parameters in CFCCIF-QESA feature set are optimized on ASVSpooF 2017 V2.0 dataset and used on the other two datasets (namely, ASVSpooF 2019 PA, and VSDC) and not vice-versa.
- If there is low frequency environmental noise present, the performance of the VLD system is expected to degrade.
- For the VLD task, we have assumed the distance of the speaker from the 7th microphone to be fixed as 5 cm, thereby we have not considered the distance variability caused due to the head movement of the speaker.
- Till now, VLD is analyzed predominantly w.r.t. replay attacks, and therefore, the scope of VLD in other spoofing techniques, such as VC and SS, remains to be explored.

8.3 Future Research Directions

- The parameters of the proposed feature sets vary w.r.t. every dataset. Thus, there is a need to come up with an approach that gives optimized parameters across all the datasets used in this study.
- Apart from cochlear filter-based features, additional auditory features such as Gammatone Frequency Cepstral Coefficients (GFCC) [309] can be explored for replay SSD.

- The proposition of u-vector is based on second-order moments of speech signal. To that effect, a study on the higher-order moments of speech signal could be performed in future to detect spoofed speech.
- For the VLD task, the distance variability caused due to the movement of the speaker’s head can be considered by using source localization techniques.
- The effect of gender on target selection on the basis of the vulnerability of speakers can be studied.
- Medical data is sensitive and difficult to collect. Therefore, the problems of infant cry classification and dysarthric severity-level classification suffer from limited data collected in restrained recording conditions and environments. To that effect, data augmentation techniques can be explored, apart from tempo, volume, and speed perturbation.

8.4 Open Research Problems

- It can be observed that even though the formal research in the anti-spoofing field started nearly a decade ago, however, still today there is no known statistically meaningful corpora for identical twins or professional impersonation; indicating the significant challenge associated with development of speech corpora for these two kinds of 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.
- In the context of CFCCIF-QESA feature set, the development of IF estimation algorithm that alleviates all the difficulties associated with the definition of IF remains an open research problem.
- In order to fool a VLD system, any low frequency noise can be artificially induced in the spoof speech utterance, so that it gets detected as pop noise. Thus, the current VLD system is not a sufficient and standalone system to detect live speech in realistic scenarios.
- Regarding anti-spoofing research, a number of corpora are publicly available in the literature, including the ASVSpooF 2015, 2017, 2019, and 2021 datasets. These common datasets, however, are only available in a predetermined set of configurations for data collection. Furthermore,

- some presumptions are used in the preparation of datasets. Such presumptions prevent us from creating generalized anti-spoofing systems that are practical for use in the actual world.
- There is no dataset that intends to build CMs for multiple spoofing attacks, or even all of them. Therefore, there is still a long way to go until generalised CMs that are appropriate for SSD deployment in the real world are developed.
 - The attacking approaches discussed in Chapter 6, are few of the many possibilities of attacks and are yet to be supported with experimental analysis.

Appendix A. Analytic Signal

The Hilbert transform of a real-valued signal $x_R(t)$ is defined as

$$\hat{x}_R(t) = \frac{1}{\pi} p.v. \int_{-\infty}^{+\infty} \frac{x_R(\tau)}{t - \tau} d\tau \quad (\text{A.1})$$

provided this integral exists as a principal value (p.v.). Furthermore, in the frequency domain, the Hilbert transformer has the frequency response of $-j \cdot \text{sgn}(\omega)$, where $\text{sgn}(\omega)$ is the signum function in the frequency domain. An analytic signal is computed using the Hilbert transform as

$$x_a(t) = x_R(t) + j\hat{x}_R(t), \quad (\text{A.2})$$

where $\hat{x}_R(t)$ denotes the Hilbert transform of $x_R(t)$. On solving the eq. (A.2) in the frequency domain, we get,

$$\begin{aligned} X_a(\omega) &= X(\omega) + j\mathcal{F}\{\hat{x}_R(t)\} \\ &= X(\omega) + j(-jX(\omega)), \quad \omega \geq 0 \\ &= X(\omega) + j(jX(\omega)), \quad \omega < 0 \end{aligned} \quad (\text{A.3})$$

On further solving, eq. (A.3) becomes

$$X_a(\omega) = \begin{cases} 2X(\omega) & \text{if } \omega \geq 0 \\ 0 & \text{if } \omega < 0 \end{cases} \quad (\text{A.4})$$

Therefore, it can be said that an analytic signal is causal in the frequency domain.

Appendix B. Teager Energy Operator (TEO)

Let us consider a discrete-time signal $x(n) = A \cos(\omega n + \theta)$, which represents the simple harmonic motion corresponding to $x(t) = A \cos(\Omega t + \phi)$ created by the mass-spring system as shown in Figure B.1. Furthermore,

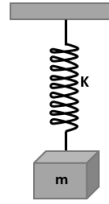


Figure B.1: A mass-spring system.

the immediate past samples of the signal can be expressed as $x(n-1) = A \cos(\omega(n-1) + \theta)$, and the immediate future samples can be expressed as $x(n+1) = A \cos(\omega(n+1) + \theta)$. Solving $x(n-1)x(n+1)$, using the trigonometric identity

$$\cos(c+d)\cos(c-d) = \frac{1}{2} [\cos(2c) + \cos(2d)], \quad (\text{B.1})$$

we get,

$$x(n+1)x(n-1) = \frac{A^2}{2} [\cos(2\omega n + 2\theta) + \cos(2\omega)]. \quad (\text{B.2})$$

On further solving using $\cos(2c) = 2\cos^2(c) - 1 = 1 - 2\sin^2(c)$, we get,

$$x(n+1)x(n-1) = A^2 \cos^2(\omega n + \theta) - A^2 \sin^2(\omega) = x^2(n) - A^2 \sin^2(\omega). \quad (\text{B.3})$$

Hence,

$$A^2 \sin^2(\omega) = x^2(n) - x(n-1) \cdot x(n+1) = \psi_R\{x(n)\}, \quad (\text{B.4})$$

where $\psi_R\{x(n)\}$ is the Teager Energy Operator (TEO) on $x(n)$. For $\omega < \pi/2$, $\sin(\omega) \approx \omega$, and hence, we can say $x^2(n) - x(n-1) \cdot x(n+1)$, which is

analogous to the energy given (in the physical sense) by

$$E = \frac{1}{2}mA^2\Omega^2. \tag{B.5}$$

Appendix C. Modelling Speech as an AM-FM Signal

Consider the discrete-time Frequency Modulated (FM) signal

$$f(n) = \cos(\phi(n)) = \cos[\omega_c n + \beta \sin(\omega_f n) + \theta], \quad (\text{C.1})$$

where $\beta = \omega_m/\omega_f$, and the IF of the signal $f(n)$ is $\omega_i(n) = d\phi(n)/dn = \omega_c + \omega_m \cos(\omega_f n)$. In eq. (C.1), let $\omega_c n + \beta \sin(\omega_f n) = C$, and let $\omega_c + \beta \sin(\omega_f) \cos(\omega_f n)$, then $f(n+1)f(n-1) = (\cos(2C) + \cos(2D))/2 = \cos^2(C) - \sin^2(D)$. If ω_f is sufficiently small such that $\cos(\omega_f) \approx 1$ and $\sin(\omega_f) \approx \omega_f$, then $\cos(C) \approx x(n)$, $D \approx \omega_i(n)$, and the applying TEO on $f(n)$, we get [172],

$$\psi[\cos(\phi(n))] \approx \sin^2[\omega_c + \omega_m \cos(\omega_f n)]. \quad (\text{C.2})$$

Given that steady-state vowels in a speech signal have time-varying formants and amplitudes, a single speech resonance can be modelled by a damped AM-FM model as:

$$x(n) = Ar^n \cos(\omega_a n) \cos[\omega_c n + \beta \sin(\omega_f n) + \theta], \quad (\text{C.3})$$

where ω_c is the center frequency of the formant, and the IF $\omega_i(n) = \omega_c + \omega_m \cos(\omega_f n)$ models the time-varying formant, and the amount of FM is controlled by $\omega_m = \beta\omega_f$. The amplitude variations are tracked by the AM envelope $|\cos(\omega_a n)|$, and the rate of energy dissipation is denoted by r . Using eq. (C.1) and eq. (C.2), we get,

$$\sqrt{\psi[x(n)]} \approx |Ar^n \cos(\omega_a n) \sin(\omega_c + \omega_m \cos(\omega_f n))|. \quad (\text{C.4})$$

This approximation is based on the assumption that ω_f is small, and $\omega_a \ll \omega_c$ [172]. Thus, $\sqrt{\psi[x(n)]}$ is a product of the envelope and the (sine of the) IF

of the resonance. Therefore, this class of signals, i.e., AM-FM signals serve as a model of energy pulses observed in actual speech signals [172].

Appendix D. IF Estimation using ESA

In [167], three DESA algorithms are mentioned, namely, DESA-1a, DESA-1, and DESA-2. In DESA-1a, '1' implies the derivative approximation in TEO with single sample difference, and a implies 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 work, DESA-1a is utilized for energy separation [167,173].

Let us consider a discrete-time AM-FM signal $y(n) = a(n)\cos(\phi(n))$, whose instantaneous frequency (IF) $\omega_i(n)$ is a finite sum of cosines. Its backward difference is given as:

$$\begin{aligned} s(n) &= y(n) - y(n-1), \\ &= a(n)c(n) + [a(n) - a(n-1)]\cos(\phi(n-1)), \\ &= D(n) + E(n) \end{aligned} \tag{D.1}$$

where

$$D(n) = a(n)c(n), \tag{D.2}$$

$$E(n) = a(n)c(n) + [a(n) - a(n-1)]\cos(\phi(n-1)). \tag{D.3}$$

Furthermore,

$$\begin{aligned} c(n) &= \cos(\phi(n)) - \cos(\phi(n-1)), \\ &= 2\sin\left(\frac{\phi(n) + \phi(n-1)}{2}\right)\sin\left(\frac{\phi(n-1) - \phi(n)}{2}\right). \end{aligned} \tag{D.4}$$

Using general approximations results for $\phi(n)$:

$$\phi(k) + \phi(m) \approx 2\phi\left(\frac{k+m}{2}\right) \quad \text{if } \omega_f|k-m| \ll 2, \quad (\text{D.5})$$

$$\phi(k) - \phi(m) \approx (k-m)\omega_i\frac{k+m}{2} \quad \text{if } \omega_f|k-m| \ll 2. \quad (\text{D.6})$$

If $\omega_f \ll 1$, we obtain from eq. (D.4):

$$c(n) \approx -2\sin(\omega_i(n-0.5)/2)\sin(\phi(n-0.5)). \quad (\text{D.7})$$

Furthermore, according to Lemma 2 in [167], the order of magnitude of E and D in eq. (D.1) are:

$$\begin{aligned} D_{max} &\approx 2\sin(\omega_i/2)_{max}a_{max}, \\ E_{max} &\approx 2\sin(\omega_a/2)a_{max}. \end{aligned} \quad (\text{D.8})$$

If $a(n)$ is band-limited, then the order of magnitude of D is much larger than that of E . Thus, ignoring E , we get,

$$y(n) \approx -2a(n)\sin(\omega_i(n-0.5)/2)\sin(\phi(n-0.5)). \quad (\text{D.9})$$

Considering the first-order approximation for standard series expansions for $\sin(\cdot)$ and $\cos(\cdot)$ on band-limited signal:

$$\psi\{s(n)\} \approx 4a^2(n)\sin^2(\omega_i(n-0.5)/2)\sin^2(\omega_i(n-0.5)). \quad (\text{D.10})$$

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:

$$|a(n)| \approx \sqrt{\frac{2\psi\{y(n)\}}{1 - \left(1 - \frac{\psi\{y(n)-y(n-1)\}}{2\psi\{y(n)\}}\right)'}} \quad (\text{D.11})$$

$$\omega_{if}(n) = \arccos \left[1 - \frac{\psi\{y(n)-y(n-1)\}}{2\psi\{y(n)\}} \right]. \quad (\text{D.12})$$

Appendix E. Heisenberg's Uncertainty Principle in Signal Processing Framework

Heisenberg's Uncertainty Principle in Signal Processing Framework

The variance in time-domain (denoted by σ_t^2), and the frequency-domain (denoted by σ_ω^2) of a window $x(t) \in L^2(\mathbb{R})$ are related by the following inequality:

$$\sigma_t^2 \cdot \sigma_\omega^2 \geq \frac{1}{4}. \quad (\text{E.1})$$

The inequality in eq. (E.1) becomes an *equality* if and only if $x(t)$ is a Gaussian, or more generally any Gabor atom.

Proof: This proof assumes fast decay of the window function $x(t) \in L^2(\mathbb{R})$, however, this theorem is valid for any $x(t) \in L^2(\mathbb{R})$ [6]. Let us consider the integral I as,

$$I = \int_{t \in \mathbb{R}} (t \cdot x(t))(x'(t)) dt = \langle tx(t), x'(t) \rangle. \quad (\text{E.2})$$

Using Cauchy-Schwartz inequality, we have

$$\left| \int_{t \in \mathbb{R}} tx(t)x'(t) dt \right| \leq \left[\int_{-\infty}^{+\infty} |tx(t)|^2 dt \right]^{\frac{1}{2}} \times \left[\int_{-\infty}^{+\infty} |x'(t)|^2 dt \right]^{\frac{1}{2}}. \quad (\text{E.3})$$

Since the window $x(t)$ has unit norm, i.e., $\|f(t)\| = 1$, we have,

$$\int_{-\infty}^{+\infty} t^2 |x(t)|^2 dt = \sigma_t^2. \quad (\text{E.4})$$

Furthermore, using the Plancherel's theorem, we get,

$$\int_{-\infty}^{+\infty} |x'(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\mathcal{F}(x'(t))|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \omega^2 |X(\omega)|^2 d\omega. \quad (\text{E.5})$$

Using integration by parts in eq. (E.2), we get,

$$I = \left[\left(\frac{t}{2} \right) \int \frac{d}{dt} x^2(t) dt \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \left(\frac{d}{dt} \left(\frac{t}{2} \right) \cdot \int \frac{d}{dt} x^2(t) dt \right) dt \} = \frac{-1}{2}. \quad (\text{E.6})$$

$$|I|^2 = 1/4 \Rightarrow \sigma_t^2 \cdot \sigma_\omega^2 \geq \frac{1}{4}. \quad (\text{E.7})$$

Cauchy-Schwartz's inequality becomes equality for collinear vectors, i.e., $b = -ka$, for $k > 0$.

$$\therefore x'(t) = -ktx(t), \quad (\text{E.8})$$

where k is scalar. Solving the differential equation, we get,

$$\int \frac{dx(t)}{dt} = \int -ktdt, \quad (\text{E.9})$$

$$\log_e x(t) = -kt^2, \quad (\text{E.10})$$

$$\therefore x(t) = e^{-kt^2}. \quad (\text{E.11})$$

It can be observed that eq. (E.11) represents a Gaussian window. Thus, this result proves that the lower bound on the area of Heisenberg's box (i.e., $\sigma_t^2 \cdot \sigma_\omega^2$) is achieved for a Gaussian window function, or more generally, any Gabor atom.

Appendix F. Energy Conservation of Time-Frequency Transforms

Energy Conservation in STFT

If a signal $x(t) \in L^2(\mathbb{R})$, the energy conservation in STFT is given by [6]:

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sx(u, \zeta)|^2 d\zeta du, \quad (\text{F.1})$$

where $Sx(u, \zeta)$ is the STFT computed at the time-frequency indices u and ζ , which vary across \mathbb{R} , covering the entire time-frequency plane. The signal reconstruction is given by [6]:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Sx(u, \zeta) g(t - u) e^{i\zeta t} d\zeta du. \quad (\text{F.2})$$

Applying Parseval's formula to eq. (F.2) w.r.t. to the integration in u , we get,

$$Sx(u, \zeta) = e^{-iu\zeta} x * g_\zeta(u), \quad (\text{F.3})$$

where $g_\zeta(t) = g(t)e^{i\zeta t}$ and $*$ indicates the convolution operator. Therefore, the Fourier transform of $Sx(u, \zeta)$ is $X(\omega_\zeta)G(\omega)$. Applying Plancherel's formula to eq. (F.1) gives

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sx(u, \zeta)|^2 dud\zeta = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(\omega + \zeta)G(\omega)|^2 d\omega d\zeta. \quad (\text{F.4})$$

Lastly, the Fubini theorem and the Plancherel formula lead to $\frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(\omega + \zeta)|^2 d\zeta = \|x\|^2$, which proves the energy conservation of STFT as shown in eq. (F.1), which justifies that the time-frequency sum of STFT is equal to the overall energy of the signal. Thus, this energy conservation gives a guarantee of the existence of a time-frequency to detect the presence of pop noise.

Energy Conservation in AWT

The inverse wavelet formula reconstructs the analytic part of a signal x as:

$$x_a(t) = \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^\infty Wx_a(u, s) \psi_s(t - u) \frac{ds}{s^2} du, \quad (\text{F.5})$$

where

$$C_\psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty \Rightarrow \int_{-\infty}^\infty \psi(t) dt = 0. \quad (\text{F.6})$$

Applying the Plancherel formula for energy conservation for the analytic part of x_a given by [6]:

$$\int_{-\infty}^{+\infty} |x_a(t)|^2 dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |W_a x(u, s)|^2 du \frac{ds}{s^2}. \quad (\text{F.7})$$

Since $Wx_a(u, s)$ is $2Wx(u, s)$, and $\|x_a\|^2$ is $2\|x\|^2$, the variable change of ζ to $\frac{1}{s}$ is done in energy conservation expression [6], which leads to

$$\|x\|^2 = \frac{2}{C_\psi} \int_0^\infty \int_{-\infty}^\infty P_w x(u, \zeta) du d\zeta. \quad (\text{F.8})$$

It again reinforces the notion that a scalogram represents time-frequency *energy density*.

Bibliography

- [1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.
- [2] P. Gupta and H. A. Patil, "Voice biometrics: Attackers perspective," in *Voice Biometrics: Technology, Trust and Security*, Carmen Gracia-Mateo and Gerard Chollet (Eds.), IET (UK), pp. 39-65, 2021, {Last Accessed: 2023-03-01}. [Online]. Available: https://digital-library.theiet.org/content/books/10.1049/pbse012e_ch3
- [3] I. Goodfellow, Y. Bengio, and A. Courville, "Convolutional networks," in *Deep Learning*. MIT Press Cambridge, MA, USA, 2016, vol. 2016, pp. 330–372.
- [4] P. Gupta, P. K. Chodingala, and H. A. Patil, "Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components," *Computer Speech & Language*, vol. 77, p. 101423, 2023.
- [5] H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 726–730.
- [6] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Elsevier, 1999.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, pp. 1097–1105, 03-06 Dec., 2012, Nevada, USA.
- [8] J. M. Lilly and S. C. Olhede, "Generalized Morse wavelets as a superfamily of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6036–6041, 2012.
- [9] Lilly, Jonathan M and Olhede, Sofia C, "Higher-order properties of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 146–160, 2008.
- [10] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 2nd Edition, Pearson Education India, 2004.
- [11] P. Gupta, S. Singh, G. P. Prajapati, and H. A. Patil, *Voice Privacy in Biometrics*. Cham: Springer International Publishing, 2023, pp. 1–29.
- [12] M. R. Portnoff, "A Quasi-One-Dimensional Digital Simulation for the Time-Varying Vocal Tract." Ph.D. dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology, USA, 1973.
- [13] "The Voice Privacy 2020 Challenge Evaluation Plan, {Last Accessed: 2021-03-15}." [Online]. Available: <https://www.voiceprivacychallenge.org>

- [14] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. L. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the voice privacy initiative," in *INTERSPEECH*, Shanghai, China, 24-28 October, 2020.
- [15] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, 7-10 December, 2020, pp. 543–549.
- [16] H. A. Patil, P. Dutta, and T. Basu, "On the Investigation of Spectral Resolution Problem for Identification of Female Speakers in Bengali," in *2006 IEEE International Conference on Industrial Technology (ICIT)*. Mumbai, India: IEEE, 2006, pp. 375–380.
- [17] D. R. Stinson and M. Paterson, *Cryptography: Theory and Practice*. 4th Edition, CRC press, 2018.
- [18] A. Kachhi, A. Therattil, P. Gupta, and H. A. Patil, "Continuous wavelet transform for severity-level classification of dysarthria," in *Speech and Computer, LNCS, Springer*, S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Eds. LNCS, Sprinkler, 2022, pp. 312–324.
- [19] 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, 2015, pp. 2037–2041.
- [20] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements," *Odyssey- The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France*, p. 296–303, 26-29 June 2018.
- [21] 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.
- [22] 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.
- [23] 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. 2019, pp. 1008–1012.
- [24] 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. 2016, pp. 1–6.
- [25] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," in *INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 2355–2359.

- [26] 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 *INTER-SPEECH*, Shanghai, China, October 2020, pp. 1081–1085.
- [27] 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 *INTER-SPEECH, Dresden, Germany*, 2015, pp. 239–243.
- [28] Priyanka Gupta, Piyushkumar K. Chodingala, and H. A. Patil, "Morlet wavelet-based voice liveness detection using convolutional neural network," in *European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 29 Aug - 02 Sep, 2022, pp. 100–104.
- [29] Priyanka Gupta, Piyushkumar K. Chodingala, and Hemant A. Patil, "Morse wavelet features for pop noise detection," in *IEEE International Conference on Signal Processing and Communication (SPCOM)*, IISc Bengaluru, India, 11-15 July, 2022, pp. 1–5.
- [30] 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.
- [31] K. M. Malik, H. Malik, and R. Baumann, "Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 523–528.
- [32] 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.
- [33] 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 2004, pp. 145–148.
- [34] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [35] 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, October, 2012.
- [36] 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 2014, pp. 1–6.
- [37] A. Paul, R. K. Das, R. Sinha, and S. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *2016 International Conference on Signal Processing and Communications (SPCOM)*, IISc, Bengaluru, India, 12-15 June 2016, pp. 1–5.
- [38] G. P. Prajapati, , M. R. Kamble, and H. A. Patil, "Energy separation based features for replay spoof detection for voice assistant." *28th European Signal Processing Conference (EUSIPCO)*, pp. 386–390, 18-21 January, 2020.

- [39] 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.
- [40] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *International Symposium on Intelligent Multimedia, Video, and Speech Processing*, Hong Kong, 20-22 October, 2004, pp. 145–148.
- [41] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH*, Lyon, France, August 2013, pp. 925–929.
- [42] L. Kersta and J. Colangelo, "Spectrographic speech patterns of identical twins," *The Journal of the Acoustical Society of America (JASA)*, vol. 47, no. 1A, pp. 58–59, 1970.
- [43] H. A. Patil and K. K. Parhi, "Variable length teager energy based mel cepstral features for identification of twins," in *International conference on pattern recognition and machine intelligence (PReMI), LNCS*, Eds. Santanu Chaudhury, Sushmita Mitra, C. A. Murthy, P. S. Sastry, Sankar K. Pal. Springer, 2009, pp. 525–530.
- [44] "Hsbc reports high trust levels in biometric tech as twins spoof its voice id system," *Biometric Technology Today*, vol. 2017, no. 6, p. 12, 2017, {Last Accessed: 2021-03-15}. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0969476517301194>
- [45] E. Team, "Twins fool HSBC voice biometrics - BBC," May 2017, {last accessed: 2021-03-15}. [Online]. Available: <https://www.finextra.com/newsarticle/30594/twins-fool-hsbc-voice-biometrics--bbc>
- [46] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP)*, vol. 1, Georgia, USA, 7-10 May, 1996, pp. 373–376.
- [47] Y. Qian, F. K. Soong, and Z.-J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 280–290, 2012.
- [48] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2017.
- [49] 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.
- [50] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *INTERSPEECH*, Antwerp, Belgium, 27-31 August, 2007, {Last Accessed: 2020-09-14}. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-02157147>
- [51] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

- [52] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *INTERSPEECH*, Antwerp, Belgium, pp. 1965-1968, 27-31 August, 2007.
- [53] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary, 5-9 September, 1999.
- [54] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, Vigo, Spain, 2010, pp. 131-134.
- [55] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," *Biometrics and ID Management*, Brandenburg, Germany, pp. 274-285, 2011.
- [56] W. Zhizheng, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2037-2041.
- [57] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC Anti-Spoofing systems for the ASVspoof 2015 challenge," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, 20-25 March, 2016, pp. 5475-5479.
- [58] M. Wester, Z. Wu, and J. Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2047-2051.
- [59] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2092-2096.
- [60] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2082-2086.
- [61] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high-dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *INTERSPEECH*, Dresden, Germany, 6-10 September, 2015, pp. 2052-2056.
- [62] 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.
- [63] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 challenge." in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 7-11.
- [64] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 27-31.
- [65] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing." in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 32-36.

- [66] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.
- [67] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH, Lyon, France*, 25-28 August 2013, pp. 925–929.
- [68] Z. Wu, A. Larcher, K. A. Lee, E. S. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: The effect of text constraints," in *INTERSPEECH, Lyon, France*, 25-28 August 2013, pp. 950–954.
- [69] W. Zhizheng, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVSpooF 2015: the first automatic speaker verification spoofing and countermeasures challenge," *INTERSPEECH, Dresden, Germany*, pp. 2037–2041, 6-10 September 2015.
- [70] 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.
- [71] 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.
- [72] 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.
- [73] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "ASVSpooF 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2021, {Last Accessed: 01-03-2023}. [Online]. Available: <https://www.asvspoof.org/workshop>
- [74] 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.
- [75] 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. 2018, pp. 681–685.
- [76] 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. 2019, pp. 1033–1037.
- [77] Q. Li, "Solution for pervasive speaker recognition," SBIR Phase I Proposal, Submitted to NSF IT.F4, Li Creative Technologies, Inc., NJ, June 2003.
- [78] 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*, pp. 2062-2066, Dresden, Germany, 6-10 September, 2015.
- [79] 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. 2019, pp. 2898–2902.

- [80] 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, 2009, pp. 181–184.
- [81] 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, 2010.
- [82] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.
- [83] 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.
- [84] M. Stéphane, "Chapter 4 - Time Meets Frequency," in *A Wavelet Tour of Signal Processing*, 3rd ed. Boston: Academic Press, 2009, pp. 89–153.
- [85] Y. Nishida, T. Hori, T. Suehiro, and S. Hirai, "Monitoring of breath sound under daily environment by ceiling dome microphone," in *International Conference on Systems, Man and Cybernetics*, vol. 3. Nashville, USA: IEEE, 2000, pp. 1822–1829.
- [86] K. Vara Prasad Naraharisetti, "Enhancement of breathing signal using delay-less subband adaptive filter with HPF," in *The 10th IEEE International Symposium on Signal Processing and Information Technology*, 15-18 December, 2010, Luxor, Egypt, pp. 177–181.
- [87] 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 - May 2019, pp. 2062–2070.
- [88] 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, 2018, pp. 233–239.
- [89] G. W. Elko, J. Meyer, S. Backer, and J. Peissig, "Electronic pop protection for microphones," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 21-24 October, 2007, pp. 46–49.
- [90] Y. Hsu, "Spectrum analysis of base-line-popping noise in MR heads," *IEEE Transactions on Magnetics*, vol. 31, no. 6, pp. 2636–2638, 1995.
- [91] S. Singh, K. Khoría, 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. 2021, pp. 649–659.
- [92] Priyanka Gupta, Siddhant Gupta, and Hemant A. Patil, "Voice Liveness Detection using Bump Wavelet with CNN," in *International Conference on Pattern Recognition and Machine Intelligence (PREMI)*, ISI Kolkata, India. Springer, 15-18 December, 2021.
- [93] K. Khoría, A. T. Patil, and H. A. Patil, "On significance of constant-Q transform for pop noise detection," *Computer Speech & Language*, vol. 77, p. 101421, 2023.

- [94] C. J. Pike, "Analysis of high resolution marine seismic data using the wavelet transform," in *Wavelet Analysis and Its Applications*. Elsevier, 1994, vol. 4, pp. 183–211.
- [95] S. Mallat, S. Zhong *et al.*, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 14, no. 7, pp. 710–732, 1992.
- [96] P. L. Goupillaud, A. Grossmann, and J. Morlet, "A simplified view of the cycle-octave and voice representations of seismic signals," in *SEG Technical Program Expanded Abstracts 1984*. Society of Exploration Geophysicists, 1984, pp. 379–382.
- [97] P. Goupillaud, A. Grossmann, and J. Morlet, "Cycle-octave and related transforms in seismic signal analysis," *Geoexploration*, vol. 23, no. 1, pp. 85–102, 1984.
- [98] A. Grossmann, R. Kronland-Martinet, and J. Morlet, "Reading and understanding continuous wavelet transforms," in *Wavelets*. Springer, 1990, pp. 2–20.
- [99] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," *Journal of Sound and Vibration*, vol. 234, no. 1, pp. 135–148, 2000.
- [100] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 44–56, 2017.
- [101] 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*, Bilbao, Spain, 2016, pp. 259–263.
- [102] K. Khoría, Ankur T. Patil, and Hemant A. Patil, "Significance of Constant-Q transform for voice liveness detection," in *29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 23-27 August 2021.
- [103] Siddhant Gupta, Kuldeep Khoría, 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)*, 14-17 Dec. 2021, Tokyo, Japan.
- [104] S. Singh, K. Khoría, and H. A. Patil, "Modified group delay cepstral coefficients for voice liveness detection," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 146–150.
- [105] A. Adiga, M. Magimai, and C. S. Seelamantula, "Gammatone wavelet cepstral coefficients for robust speech recognition," in *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*. IEEE, 2013, pp. 1–4.
- [106] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, "Auditory-motivated gammatone wavelet transform," *Signal Processing*, vol. 94, pp. 608–619, 2014.
- [107] 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 2021, pp. 126–130.

- [108] P. Kocher, J. Jaffe, B. Jun, and P. Rohatgi, "Introduction to differential power analysis," *Journal of Cryptographic Engineering*, vol. 1, no. 1, pp. 5–27, 2011.
- [109] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *Annual International Cryptology Conference*. Santa Barbara, California, USA: Springer, 15-19 August, 1999, pp. 388–397.
- [110] R. Kumar, P. Jovanovic, W. Burleson, and I. Polian, "Parametric trojans for fault-injection attacks on cryptographic hardware," in *In IEEE, Workshop on Fault Diagnosis and Tolerance in Cryptography*, Busan, South Korea, 23 September, 2014, pp. 18–28.
- [111] "Document iso/iec 24745:2011, information technology- security techniques-biometric information protection," 2011, iSO/IEC JTC1 SC27 Security Techniques.
- [112] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2017.
- [113] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in asr: Reality or illusion?" *arXiv preprint arXiv:1911.04913*, 2019, {Last Accessed: 2020-08-09}.
- [114] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.
- [115] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," *arXiv preprint arXiv:1803.00860*, 2018, {Last Accessed: 2020-08-10}.
- [116] V. Vestman, B. Soomro, A. Kanervisto, V. Hautamäki, and T. Kinnunen, "Who do i sound like? showcasing speaker recognition technology by youtube voice search," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 12-17 May, 2019, pp. 5781–5785.
- [117] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.
- [118] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi, United Arab Emirates, 2-6 April, 2017, pp. 506–519.
- [119] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real Bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP), virtual*. IEEE, May 24-27, 2021, pp. 694–711.
- [120] Yee Wah Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. of International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, 20-22 Oct. 2004, pp. 145–148.

- [121] T. Kinnunen, R. G. Hautamäki, V. Vestman, and M. Sahidullah, "Can we use speaker recognition technology to attack itself? Enhancing Mimicry Attacks Using Automatic Target Speaker Selection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, May 12-17, 2019, pp. 6146–6150.
- [122] X. Tian, R. K. Das, and H. Li, "Black-box attacks on automatic speaker verification using feedback-controlled voice conversion," *Odyssey 2020 The Speaker and Language Recognition Workshop*, pp. 159-164, November 1-5, 2020, Tokyo, Japan.
- [123] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, 4-8 May, 2020, pp. 6579–6583.
- [124] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial regularization for end-to-end robust speaker verification," in *INTERSPEECH*, Graz, Austria, 15-19 September, 2019, pp. 4010–4014.
- [125] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems," *Journal of Signal Processing Systems*, pp. 1–14, vol. 93, 2021.
- [126] A. Gomez-Alanis, J. A. Gonzalez, and A. M. Peinado, "Adversarial Transformation of Spoofing Attacks for Voice Biometrics," in *Proc. INTERSPEECH 2021*, Valladolid, Spain, March 24-25, 2021, pp. 255–259, {Last Accessed: 02-04-2021}. [Online]. Available: <http://dx.doi.org/10.21437/IberSPEECH.2021-54>
- [127] M. Pal and G. Saha, "On robustness of speech based biometric systems against voice conversion attack," *Applied Soft Computing*, vol. 30, pp. 214–228, 2015.
- [128] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Information Processing Systems*, vol. 31, pp. 4485-4495, Montreal, 3-8 Dec., 2018.
- [129] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 10-15 June, 2019., pp. 5210–5219.
- [130] Y. Gao, J. Lian, B. Raj, and R. Singh, "Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems," in *IEEE Spoken Language Technology Workshop (SLT)*, Virtual Conference, 19-22 January, 2021, pp. 544–551.
- [131] 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: ACM, 2017, pp. 103–117.
- [132] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *European Symposium on Research in Computer Security*. Springer, 21-25 September, 2015, Vienna, Austria, pp. 599–621.

- [133] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, Taiwan, October 5 - 9, 2020, pp. 357–369.
- [134] E. Zetterholm, M. Blomberg, and D. Elenius, "A comparison between human perception and a speaker verification system score of a voice imitation," *evaluation*, vol. 119, no. 116.4, pp. 116–4, 2004.
- [135] 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.
- [136] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 651–655.
- [137] 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. 2015.
- [138] 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. 2019, pp. 1073–1077.
- [139] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.
- [140] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *INTERSPEECH*, Portland, OR, USA, 9-13 September 2012, pp. 1448–1451.
- [141] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.
- [142] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2011.
- [143] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.
- [144] L. Cohen, *Time-Frequency Analysis*. Prentice-Hall, 1995, vol. 778.
- [145] 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.
- [146] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

- [147] 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.
- [148] 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 1992, pp. 137–140.
- [149] 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. 2015, pp. 2996–3000.
- [150] 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>
- [151] 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}.
- [152] 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.
- [153] C. M. Bishop, *Pattern Recognition and Machine Learning*. 3rd Edition, Springer, 2006.
- [154] D. A. Reynolds, "Gaussian mixture models." *Encyclopedia of Biometrics*, vol. 741, pp. 659–663, 2009.
- [155] 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.
- [156] 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 2010, pp. 253–256.
- [157] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning (ICML)*, Haifa, Israel, 21-24 June 2010, pp. 807–814.
- [158] 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, 20-24, Aug 2017*, pp. 82–86.
- [159] 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.
- [160] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *INTERSPEECH*, Graz, Austria, 2019, pp. 1013–1017.
- [161] W. H. Kang, J. Alam, and A. Fathan, "CRIM's System Description for the ASVSpooof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 2021, pp. 100–106.

- [162] Z. Benhafid, S. A. Selouani, M. S. Yakoub, and A. Amrouche, "LARIHS AS-SERT Reassessment for Logical Access ASVspooof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 2021, pp. 94–99.
- [163] 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. 2021, pp. 83–88.
- [164] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspooof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 2021, pp. 89–93.
- [165] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspooof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Sept. 2021, pp. 61–67.
- [166] R. D. Nindrea, T. Aryandono, L. Lazuardi, and I. Dwiprahasto, "Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis," *Asian-Pacific Journal of Cancer Prevention: APJCP*, vol. 19, no. 7, p. 1747, 2018.
- [167] 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.
- [168] 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.
- [169] 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 1990, pp. 381–384.
- [170] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal: I. fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, 1992.
- [171] J. F. Kaiser, "On Teager's energy algorithm and its generalization to continuous signals," in *Proc. 4th IEEE Digital Signal Processing Workshop*. Mohonk (New Palts), NY, 1990, pp. 338–375.
- [172] 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, May 1991, pp. 421–424.
- [173] 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.
- [174] 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.
- [175] 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.

- [176] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal (BSTJ)*, vol. 27, no. 3, pp. 379–423, 1948.
- [177] T. M. Cover, *Elements of Information Theory*. 2nd Edition, John Wiley & Sons, 1999.
- [178] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.
- [179] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L² theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.
- [180] A. V. Oppenheim, *Discrete-Time Signal Processing*. 3rd Edition, Pearson Education India, 1999.
- [181] 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.
- [182] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition." in *INTERSPEECH, Lisbon, Portugal*, 4-8 September 2005, pp. 3013–3016.
- [183] M. R. Kamble, H. Tak, and H. A. Patil, "Amplitude and frequency modulation-based features for detection of replay spoof speech," *Speech Communication*, vol. 125, pp. 114–127, 2020.
- [184] R. Bellman, "Dynamic Programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.
- [185] C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition," *Pattern Recognition and Image Analysis*, vol. 24, no. 1, pp. 124–132, March 2014.
- [186] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH, Dresden, Germany*, Sept. 2015, pp. 2087–2091.
- [187] 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, 2016, pp. 249–252.
- [188] 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.
- [189] 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. 1997, pp. 1895–1898.
- [190] Y. Huang, L. He, W. Wei, W. Gale, J. Li, and Y. Gong, "Using personalized speech synthesis and neural language generator for rapid speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain*. IEEE, 2020, pp. 7399–7403.
- [191] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

- [192] J. Monteiro, I. Albuquerque, J. Alam, R. D. Hjelm, and T. Falk, "An end-to-end approach for the verification problem: learning the right distance," in *International Conference on Machine Learning (ICML), Sydney Australia, 2020*, pp. 7022–7033.
- [193] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech." in *INTERSPEECH*, 2016, pp. 1710–1714.
- [194] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. on Info Theory*, vol. 37, no. 1, pp. 145–151, 1991.
- [195] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NIPS), Montreal, Canada*, pp. 2672–2680, 8-11 December 2014.
- [196] B. S. Atal, "Automatic speaker recognition based on pitch contours," *The Journal of the Acoustical Society of America (JASA)*, vol. 52, no. 6B, pp. 1687–1697, 1972.
- [197] 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.
- [198] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782–796, 2008.
- [199] S. Cheedella S. Gupta and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.
- [200] J. Mishra, M. Singh, and D. Pati, "Processing linear prediction residual signal to counter replay attacks," in *International Conference on Signal Processing and Communications (SPCOM), IISc, Bangaluru, India, 16-19 July, 2018*, pp. 95–99.
- [201] C. Hanilçi, "Speaker verification anti-spoofing using linear prediction residual phase features," in *2017 25th European Signal Processing Conference (EU-SIPCO), Kos Island, Greece, August 8 - September 2, 2017*, pp. 96–100.
- [202] J. D. Markel and A. J. Gray, *Linear Prediction of Speech*. Springer Science & Business Media, 2013, vol. 12.
- [203] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America (JASA)*, vol. 50, no. 2B, pp. 637–655, 1971.
- [204] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [205] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [206] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

- [207] 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)*, Quebec, Canada, November 14-16, 2017, pp. 51–55.
- [208] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989.
- [209] T. Claasen and W. Mecklenbrauker, "The Wigner distribution — a tool for time-frequency signal analysis," *Philips J. Res*, vol. 35, no. 3, pp. 217–250, 1980.
- [210] B. Bouachache and P. Flandrin, "Wigner-ville analysis of time-varying signals," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Paris, France*, vol. 7. IEEE, 3-5 May, 1982, pp. 1329–1332.
- [211] Boashash, Boualem, *Time-frequency signal analysis and processing: A Comprehensive Reference*, 2nd ed. Academic Press, 2015.
- [212] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*. 2nd Edition, Prentice Hall, 1997.
- [213] P. Busch, T. Heinonen, and P. Lahti, "Heisenberg's uncertainty principle," *Physics Reports*, vol. 452, no. 6, pp. 155–176, 2007.
- [214] 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.
- [215] M. D. Beecher, "Spectrographic analysis of animal vocalizations: implications of the "uncertainty principle," *Bioacoustics*, vol. 1, no. 2-3, pp. 187–208, 1988.
- [216] B. Boashash, "Time-frequency and instantaneous frequency concepts (chapter 1)," in *Time-Frequency Signal Analysis and Processing (Second Edition)*. Oxford: Academic Press, 2016, pp. 31 – 63.
- [217] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 137, no. 4, pp. 2047–2059, 2015.
- [218] D. O'shaughnessy, *Speech communications: Human and machine (IEEE)*. Universities press, 1987.
- [219] Q. Lin, E.-E. Jan, C. Che, D.-S. Yuk, and J. Flanagan, "Selective use of the speech spectrum and a vqgmm method for speaker identification," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 4. IEEE, 1996, pp. 2415–2418.
- [220] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, no. 4, pp. 312–322, 2008.
- [221] A. Neustein and H. A. Patil, *Forensic speaker recognition*. Springer, 2012, vol. 1.
- [222] H. A. Patil, R. Acharya, A. T. Patil, and P. Gupta, "Non-cepstral uncertainty vector for replay spoofed speech detection," in *30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, Aug. 29- Sept. 02, 2022, pp. 374–378.
- [223] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

- [224] D. Vakman and L. Vainshtein, "Amplitude, phase, frequency- fundamental concepts of oscillation theory," *Soviet Physics Uspekhi*, vol. 20, no. 12, p. 1002, 1977.
- [225] J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," *The Journal of the Acoustical Society of America*, vol. 20, no. 1, pp. 42–51, 1948.
- [226] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torresani, "Asymptotic wavelet and gabor analysis: Extraction of instantaneous frequencies," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 644–664, 1992.
- [227] MATLAB Documentation, "Bump wavelet," {Last Accessed: 29-06-2021}. [Online]. Available: <https://in.mathworks.com/help/wavelet/ref/cwtf.html#buu64ch>
- [228] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.
- [229] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.
- [230] , "Zur theoria der orthogonalen funktionsysteme," *Math. Fnn*, vol. 69, pp. 331–371, 1910.
- [231] I. Daubechies, "Where do wavelets come from? A personal point of view," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 510–513, 1996.
- [232] Lilly, Jonathan M and Olhede, Sofia C, "On the analytic wavelet transform," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 4135–4156, 2010.
- [233] P. M. Morse, "Diatomic molecules according to the wave mechanics. ii. vibrational levels," *Physical Review*, vol. 34, no. 1, p. 57, 1929.
- [234] V. Olivier *et al.*, *Airy functions and applications to physics*. World Scientific, 2010.
- [235] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *submitted in Computer Speech & Language*, 2023.
- [236] 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*, Sep. 15-19, 2019, pp. 1008–1012.
- [237] 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.
- [238] D. R. C. E. Vincent, "Roomsimove, GNU public license," {Last Accessed: 24 September, 2022}. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip
- [239] 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.

- [240] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ontario, Canada: IEEE, 6-11 June 2021, pp. 6369–6373.
- [241] B. A. Malin, K. E. Emam, and C. M. O'Keefe, "Biomedical data privacy: problems, perspectives, and recent advances, BMJ Publishing Group," pp. 2–6, 2013.
- [242] B. B. Boyer, "Computerized medical records and the right to privacy: the emerging federal response," *BuFF. L. REv.*, vol. 25, p. 37, 1975.
- [243] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.
- [244] G. Fant, *Acoustic Theory of Speech Production*. 2nd Edition, Walter de Gruyter, 1970.
- [245] B. Atal and J. Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7. IEEE, 1982, pp. 614–617.
- [246] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Atlanta, Georgia, USA: IEEE, 1996, pp. 346–348.
- [247] H. Mizuno and M. Abe, "A Formant Frequency Modification Algorithm Dealing with the Pole Interaction," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 79, no. 1, pp. 46–55, 1996.
- [248] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," *Proceedings of the IEEE*, vol. 54, no. 5, pp. 720–734, May 1966.
- [249] J. Slifka and T. R. Anderson, "Speaker Modification with LPC Pole Analysis," in *1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. Detroit, Michigan, USA: IEEE, 1995, pp. 644–647.
- [250] S. McAdams, "Spectral fusion, spectral parsing, and the formation of auditory image," *Ph.D. Thesis, Department of Hearing and Speech, Stanford University, California, USA*, May, 1984.
- [251] C. Un and D. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1466–1474, 1975.
- [252] M. Schroeder and B. Atal, "Code-excited linear prediction (celp): High-quality speech at very low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, 1985, pp. 937–940.
- [253] A. V. McCree and T. P. Barnwell, "A mixed excitation lpc vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.
- [254] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Brisbane, Australia: IEEE, 19-24 April, 2015, pp. 5206–5210.
- [255] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)," 2019.

- [256] P. Gupta, G. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "System description : Design of voice privacy system using linear prediction," 2020, {Last Accessed: 15-01-2021}. [Online]. Available: <https://www.voiceprivacychallenge.org/docs/DA-IICT-Speech-Group.pdf>
- [257] Hemant A. Patil, "Speaker recognition in indian languages: A feature based approach," *Indian Institute of Technology Kharagpur (IIT-K), Department of Electrical Engineering, Ph. D Thesis*, 2005.
- [258] B. Yegnanarayana, S. M. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Transactions on speech and audio processing*, vol. 13, no. 4, pp. 575–582, 2005.
- [259] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance," National Institute of Standards and Technology (NIST), Gaithersburg Md, Tech. Rep., 1998.
- [260] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.
- [261] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Alberta, Canada, 15-20 April, 2018, pp. 5329–5333.
- [262] B. Nguyen and F. Cardinaux, "Nvc-net: End-to-end adversarial voice conversion," *arXiv e-prints*, pp. arXiv–2106, 2021, {Last Accessed: 30 March, 2022}.
- [263] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspooF 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 1st Sept. 2021, pp. 47–54.
- [264] Y. Gong and C. Poellabauer, "An overview of vulnerabilities of voice controlled systems," 1st *International workshop on Security and privacy for Internet-of-Things*, Orlando, United States, April 2018.
- [265] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *Proceedings Network and Distributed System Security Symposium*, arXiv preprint arXiv:1704.01155, 2017, {Last Accessed: 2020-05-14}.
- [266] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X. Li, "Speech sanitizer: Speech content desensitization and voice anonymization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, issue 6, pp. 1–1, 2019.
- [267] 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.
- [268] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12-17, 2019, Brighton, UK, pp. 2307–2311.

- [269] H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, and C.-C. J. Kuo, "Defakehop: A light-weight high-performance deepfake detector," *arXiv e-prints*, pp. arXiv-2103, 2021, {Last Accessed: 26-02-2022}.
- [270] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis," in *CVPR Workshops*, Long Beach California, 16-20 June, 2019.
- [271] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.
- [272] W. Stallings, *Cryptography and Network Security: Principles and Practices*. 4th Edition, Pearson Education India, 2006.
- [273] R. L. Rivest, A. Shamir, and L. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.
- [274] X. Bai, L. Jiang, X. Liu, and J. Tan, "Rsa Encryption/Decryption Implementation Based on Zedboard," in *International Conference on Trustworthy Computing and Services*. Springer, 2014, pp. 114–121.
- [275] J. D. Dixon, "The Number of Steps in the Euclidean Algorithm," *Journal of Number Theory*, vol. 2, no. 4, pp. 414–422, 1970.
- [276] C. Gentry and D. Boneh, *A Fully Homomorphic Encryption Scheme*. Stanford University, 2009, vol. 20, no. 9.
- [277] R. Nara, K. Satoh, M. Yanagisawa, T. Ohtsuki, and N. Togawa, "Scan-based Side-Channel Attack Against RSA Cryptosystems Using Scan Signatures," *IEICE transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 93, no. 12, pp. 2481–2489, 2010.
- [278] S. A. Reijneveld, E. Brugman, and R. A. Hirasing, "Excessive infant crying: the impact of varying definitions," *Pediatrics*, vol. 108, no. 4, pp. 893–897, 2001.
- [279] 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, 2017.
- [280] 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. 2019, pp. 3053–3057.
- [281] 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. 2005, pp. 105–108.
- [282] 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.
- [283] Q. Xie, R. K. Ward, and C. A. Laszlo, "Determining normal infants' level-of-distress from cry sounds," in *Proceedings of Canadian Conference on Electrical and Computer Engineering*, Vancouver, BC, Canada, 1993, pp. 1094–1096.

- [284] H. A. Patil, "Cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry," in *A. Newstein (Ed.) Advances in Speech Recognition*, Springer, 2010, pp. 323–348.
- [285] L. Armbrüster and et. al., "Musical intervals in infants' spontaneous crying over the first 4 months of life," *Folia Phoniatrica et Logopaedica*, vol. 73, no. 5, pp. 401–412, 2021.
- [286] 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)*, Singapore, 2022, pp. 7392–7396.
- [287] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961–1005, 1990.
- [288] P. Gupta, P. K. Chodingala, and H. A. Patil, "Morse wavelet features for pop noise detection," in *International Conference on Signal Processing and Communications (SPCOM)*, IISc Bangalore, India, 2022, pp. 1–5.
- [289] J. Chunyan, M. Chen, L. Bin, and Y. Pan, "Infant cry classification with graph convolutional networks," in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, 2021, pp. 322–327.
- [290] C. Ji, Y. Jiao, M. Chen, and Y. Pan, "Infant cry classification based-on feature fusion and mel-spectrogram decomposition with cnns," in *International Conference on AI and Mobile Services*. Springer, 2022, pp. 126–134.
- [291] H.-N. Ting, Y.-M. Choo, and A. A. Kamar, "Classification of asphyxia infant cry using hybrid speech features and deep learning models," *Expert Systems with Applications*, vol. 208, p. 118064, 2022.
- [292] M. Hariharan, S. Yaacob, and S. A. Awang, "Pathological infant cry analysis using wavelet packet transform and probabilistic neural network," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15 377–15 382, 2011.
- [293] 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.
- [294] Neeharika Buddha and Hemant A. Patil, "Corpora for analysis of infant cry," O-COCOSDA, Vietnam 2007.
- [295] A. Chittora and H. A. Patil, "Data collection of infant cries for research and analysis," *Journal of Voice*, vol. 31, no. 2, pp. 252–e15, 2017.
- [296] H. F. Alaie, L. Abou-Abbas, and C. Tadj, "Cry-based infant pathology classification using GMMs," *Speech Communication*, vol. 77, pp. 28–52, 2016.
- [297] 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.
- [298] P. Lieberman, "Primate vocalizations and human linguistic ability," *The Journal of the Acoustical Society of America (JASA)*, vol. 44, no. 6, pp. 1574–1584, 1968.
- [299] 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.
- [300] C. Mackenzie and A. Lowit, "Behavioural intervention effects in dysarthria following stroke: communication effectiveness, intelligibility and dysarthria impact," *International Journal of Language & Communication Disorders*, vol. 42, no. 2, pp. 131–153, 2007.
- [301] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research (JSLHR)*, vol. 12, no. 2, pp. 246–269, 1969.
- [302] Siddhant Gupta, Ankur T. Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A. Patil, and Rodrigo Capobianco Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, 2021.
- [303] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18 183–18 194, 2021.
- [304] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 2021*, pp. 116–120.
- [305] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," in *INTERSPEECH, Stockholm, Sweden, 2017*, pp. 3127–31.
- [306] 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.
- [307] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the CUHK dysarthric speech recognition system for the UA speech corpus," in *INTERSPEECH, Hyderabad, India, 2018*, pp. 2938–2942.
- [308] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*. Springer, 2013, pp. 237–280.
- [309] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, "Auditory-motivated gammatone wavelet transform," *Signal Processing*, vol. 94, pp. 608–619, 2014.

List of Publications from the Thesis

Journal Papers

1. **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil "Replay Spoof Detection Using Energy Separation Based Instantaneous Frequency Estimation from Quadrature and In-Phase Components", in *Computer, Speech & Language*, Elsevier, vol. 77 (2023), pp. 101423.
2. **Priyanka Gupta**, and Hemant A. Patil, "Morse Wavelet Transform-based Features for Voice Liveness Detection", in *Computer, Speech & Language*, Elsevier, vol. 84, 2024.
3. **Priyanka Gupta**, Hemant A. Patil, and Rodrigo Capobianco Guido "On Vulnerability Issues in Automatic Speaker Verification (ASV) Systems", accepted under minor revision in *EURASIP Journal on Audio, Speech, and Music (JASM) Processing, Special Issue on Security & Privacy in Speech Communication*, 2023.

Chapters in Edited Books

1. **Priyanka Gupta**, Rajul Acharya, Ankur Patil and Hemant A. Patil, "On the Asymptotic Behaviour of the Speech Signal", in *25th International Conference on Speech and Computer (SPECOM)*, IIT Dharwad, India, 2023.
2. **Priyanka Gupta**, and Hemant A. Patil, "Significance of Distance on Pop Noise for Voice Liveness Detection," in *International Conference on Speech and Computer (SPECOM)*, *Lecture Notes in Computer Science (LNCS)*, vol 13721, pp. 226-237, 2022, Springer.
3. Aastha Kachhi, Anand Therattil, **Priyanka Gupta**, and Hemant A. Patil, "Continuous Wavelet Transform for Severity-Level Classification of Dysarthria," in *International Conference on Speech and Computer (SPECOM)*, Eds. S. R. Mahadeva Prasanna et.al, *Lecture Notes in Computer Science (LNCS)*, Springer, vol 13721, pp. 312–324, 2022.
4. **Priyanka Gupta**, Siddhant Gupta and Hemant A. Patil, "Voice Liveness Detection using Bump Wavelet with CNN," in *International Con-*

- ference on Pattern Recognition and Machine Intelligence (PReMI), Lecture Notes in Computer Science (LNCS), 2021, Springer.
5. **Priyanka Gupta**, and Hemant A. Patil, "Voice Biometrics: Attacker's Perspective," Gerard Chollet, and Carmen Garcia Mateo (Eds.) in Voice Biometrics: Technology, trust and security, Institution of Engineering and Technology (IET), 2021.
 6. **Priyanka Gupta**, Shrishti Singh, Gauri P. Prajapati and Hemant A. Patil, "Voice Privacy in Biometrics," in Biomedical Signal and Image Processing with Artificial Intelligence, Eds. Chirag Paunwala et al, 2023, pp. 1-29.

Conference Papers

1. **Priyanka Gupta**, Aastha Kachhi, and Hemant A. Patil, "Classification of Normal vs. Pathological Infant Cries Using Morse Wavelets", in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Taipei, Taiwan, 31 October-3 November 2023.
2. **Priyanka Gupta**, Piyushkumar K. Chodingala and Hemant A. Patil "Relevance of Quadrature Phase For Replay Detection in Voice Assistants (VAs)" in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Taipei, Taiwan, 31 October-3 November 2023.
3. Siddharth Rathod, Aastha Kachhi, **Priyanka Gupta**, and Hemant A. Patil, " Cochlear Filter-Based Cepstral Features for Dysarthric Severity-Level Classification ", in 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 04 -08 Sept., 2023.
4. **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil, "Significance of Quadrature and In-Phase Components for Synthetic Spoofed Speech Detection", in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, pp. 1252-1258, Nov. 7-10, 2022.
5. **Priyanka Gupta**, and Hemant A. Patil, "Effect of Speaker-Microphone Proximity on Pop Noise: Continuous Wavelet Transform-Based Approach" in the 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, Dec. 11-14, 2022.
6. **Priyanka Gupta**, and Hemant A. Patil "Linear Frequency Residual Cepstral Features for Replay Spoof Detection on ASVspoof 2019", in 30th

- European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, pp. 349-353, 29 Aug. -02 Sept., 2022.
7. **Priyanka Gupta**, Piyushkumar K. Chodingala and Hemant A. Patil "Energy Separation Based Instantaneous Frequency Estimation from Quadrature and In-Phase Components for Replay Spoof Detection", in 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, pp. 369-373, 29 Aug. -02 Sept., 2022.
 8. **Priyanka Gupta**, Piyushkumar K. Chodingala and Hemant A. Patil "Morlet Wavelet-Based Voice Liveness Detection using Convolutional Neural Network", in 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, pp. 100-104, 29 Aug. -02 Sept., 2022.
 9. Hemant A. Patil, Rajul Acharya, Ankur T, Patil, and **Priyanka Gupta**, "Non-Cepstral Uncertainty Vector for Replay Spoofed Speech Detection", in 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, pp. 374-378, 29 Aug. -02 Sept., 2022.
 10. Aastha Kachhi, **Priyanka Gupta**, and Hemant A. Patil "Features Motivated From Uncertainty Principle for Classification of Normal vs. Pathological Infant Cry", in 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, pp. 1253-1257, 29 Aug. -02 Sept., 2022.
 11. Anand Therattil, **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil, "Teager Energy Based-Detection of One-point and Two-point Replay Attacks: Towards Cross-Database Generalization," in Proc. The Speaker and Language Recognition Workshop (Odyssey 2022), June 28 - July 01, 2022.
 12. **Priyanka Gupta**, Piyushkumar K. Chodingala, and Hemant A. Patil, "Morse Wavelet Features for Pop Noise Detection," in 2022 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, July 11-15, 2022, pp. 1-5. **Best Paper Award Finalist**
 13. **Priyanka Gupta**, Gauri P. Prajapati, Shrishti Singh, Madhu R. Kamble and Hemant A. Patil, "Design of Voice Privacy System using Linear Prediction," in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, December 7-10, 2020, pp. 543-549.

Brief Biography



Priyanka Gupta received her B.Tech. degree in Electronics and Communication Engineering (ECE) from S.P.S.U., Udaipur, Rajasthan, India in 2014. She qualified the all-India GATE exam in 2016, and completed her M. Tech. in VLSI from The LNM Institute of Information and Communication Technology (LNMIIT), Jaipur, Rajasthan, India in 2017, under the supervision of Prof. (Dr.) Kusum Lata. During her M. Tech., she was a member of the VESD research group at LNMIIT, and her research was based on the implementation of cryptographic algorithms on Field Programmable Gate Arrays (FPGAs).

She joined DA-IICT Gandhinagar, India as a doctoral student in July 2017. After some research experience in the cryptography domain, she joined the Speech Research Lab at DA-IICT in January 2020. Currently, she is a doctoral student under the supervision of Prof. (Dr.) Hemant A. Patil at DA-IICT, Gandhinagar, India. Her primary research is focused on anti-spoofing for voice biometric systems. To that effect, she has worked in specialized areas of feature engineering for spoofed speech and voice liveness detection, design of voice privacy systems, and attacker's perspective. Furthermore, she has also contributed to the areas of pathological infant cry detection and dysarthric severity-level classification. Furthermore, she has delivered a talk jointly with Prof. Hemant A. Patil on the attacker's perspective at

Nirma University, Ahmedabad, India. She has several research publications in the form of conference papers, book chapters, and journal papers. She has attended and presented her research work at various conferences such as ANTS 2019, APSIPA-ASC 2020, PReMI 2021, SPCOM 2022, EUSIPCO 2022, APSIPSA-ASC 2022, and ISCSLP 2022. She was also the finalist for the best student paper award at SPCOM 2022 at IISc Bengaluru, for her paper titled 'Morse Wavelet Features for Pop Noise Detection'. She was also the Session Chair for one of the sessions in EUSIPCO 2022, and also received EURASIP student travel grant to present her papers in EUSIPCO 2022. She is also a reviewer for ICASSP 2023, and EURASIP Journal on Audio, Speech, and Music Processing.

At DA-IICT, she was a teaching assistant for various courses at bachelor as well as master levels, for courses such as Digital System Design, Operating Systems, Cryptography, Basic Electronics, Analog Communication & Transmission Line Theory, and Signals & Systems.