Replay Spoof Detection using Handcrafted Features

by

PRASAD ANIL TAPKIR
201611027

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

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

July, 2018

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

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

Prasad Anil Tapkir

Certificate

This is to certify that the thesis work entitled "REPLAY SPOOF DETECTION USING HANDCRAFTED FEATURES" has been carried out by PRASAD ANIL TAPKIR for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my supervision.

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

Acknowledgments

No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude. I avail this opportunity to thank all the well-wishers for their continuous and excellent support to carry out this work.

First of all, I would like to thank my supervisor, Prof. Hemant A. Patil. It cannot be argued that he is the most important and influential person to carry out this work, whose constant motivation, support and suggestions made the completion of the thesis. I would also like to thank my committee members: Prof. Suman K. Mitra and Prof. Manish K. Gupta for their insightful comments and encouragement during presentations. Foremost, I would like to thank Speech Research Lab members: Ankur Patil, Madhu Kamble, Nirmesh Shah, Hardik Sailor, Maulik Madhavi, Neil Shah, Hemlata Tak, Sreeraj R. and Srinivas Kantheti for their suggestions and help during various stages of the thesis.

My special gratitude towards my friends especially, Neil Shah and Jeni Shah for their continuous support and motivation during the tough times. I avail this opportunity to express my gratitude to my family, my parents, my sisters. I would like to thank them for their love and support to achieve my goals. Above all, I am thankful to Almighty for blessing me with good health and mental strength.

Contents

Abstract

In the past few years, there has been noteworthy demand in the use of Automatic Speaker Verification (ASV) system for numerous applications. The increased use of the ASV systems for voice biometrics purpose comes with the threat of spoofing attacks. The ASV systems are vulnerable to five types of spoofing attacks, namely, impersonation, Voice Conversion (VC), Speech Synthesis (SS), twins and replay. Among which, replay possess a greater threat to the ASV system than any other spoofing attacks, as it neither require any specific expertise nor a sophisticated equipment. Replay attacks require low efforts and most accessible attacks. The replay speech can be modeled as a convolution of the genuine speech with the impulse response of microphone, multimedia speaker, recording environment and playback environment. The detection difficulty of replay attacks increases with a high quality intermediate devices, clean recording and playback environment. In this thesis, we have propose three novel handcrafted cepstral feature sets for replay spoof detection task, namely, Magnitude-based Spectral Root Cepstral Coefficients (MSRCC), Phase-based Spectral Root Cepstral Coefficients (PSRCC) and Empirical Mode Decomposition Cepstral Coefficients (EMDCC). In addition, we explored the significance of Teager Energy Operator (TEO) phase feature for replay spoof detection.

The EMDCC feature set replace the filterbank structure with Empirical Mode Decomposition (EMD) technique to obtain the subband signals. The number of subbands obtained for the replay speech signal using EMD is more as compared to the genuine speech signal. The MSRCC and PSRCC feature sets are extracted using spectral root cepstrum of speech signal. The spectral root cepstrum spreads the effect of additional impulse responses in replay speech over entire quefrency-domain. The TEO phase feature set provides the high security information when fused with other magnitude-based features, such as Mel Frequency Cepstral Coefficients (MFCC). The experiments are performed on ASV spoof 2017 challenge database and all the systems are implemented using Gaussian Mixture Model (GMM) as a classifier. All the feature set performs better than the ASV spoof 2017 challenge baseline Constant Q Cepstral Coefficients (CQCC) system.

List of Acronyms

AM-FM	Amplitude Modulation and Frequency Modulation
ASV	Automatic Speaker Verification
BLSTM	Bidirectional Long Short Term Memory
CMN	Cepstral Mean Normalization
CMVN	Cepstral Mean Variance Normalization
CNN	Convolutional Neural Network
CQCC	Constant Q Cepstral Coefficients
CQT	Constant Q Transform
DCT	Discrete Cosine Transform
DET	Detection Error Trade-off
DNN	Deep Neural Network
EER	Equal Error Rate
EF	Epoch Features
EM	Expectation Maximization
EMD	Empirical Mode Decomposition
EMDCC	Empirical Mode Decomposition Cepstral Coefficients
ESA	Energy Separation Algorithm
FAR	False Acceptance Rate
FDNN	Fuzzy Deep Neural Network
FFT	Fast Fourier Transform
FRR	False Rejection Rate
FT	Fourier Transform
GCI	Glottal Closure Instant
GD	Group Delay
GMM	Gaussian Mixture Model
GSV-GBDT	GMM mean supervector-Gradient Boosting Decision Tree
HFCC	High Frequency Cepstral Coefficients
IF	Instantaneous Frequency
IFCC	Instantaneous Frequency Cepstral Coefficients

IMF	Intrinsic Mode Function
IMFCC	Inverted Mel Frequency Cepstral Coefficients
LCNN	Light Convolutional Neural Network
LFCC	Linear Frequency Cepstral Coefficients
LFMGDCC	Linear Frequency Modified Group Delay Cepstral Coefficients
LLR	Log-Likelihood Ratio
LP	Linear Prediction
LPCC	Linear Prediction Cepstral Coefficients
LPCCres	Linear Prediction Cepstral Coefficients residual
MFCC	Mel Frequency Cepstral Coefficients
MFMGDCC	Mel Frequency Modified Group Delay Cepstral Coefficients
MGDF	Modified Group Delay Function
MSRCC	Magnitude-based Spectral Root Cepstral Coefficients
PSRCC	Phase-based Spectral Root Cepstral Coefficients
ResNet	Residual Network
RF	Random Forest
ROC	Receiver Operating Characteristics
RNN	Recurrent Neural Network
SFF	Single Frequency Filtering
SID	Speaker Identification
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SS	Speech Synthesis
SSD	Spoof Speech Detection
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TEO	Teager Energy Operator
TTS	Text-to-Speech
VC	Voice Conversion
VESA-IFCC	Variable length Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients

List of Symbols

$r(t)$	Replay speech signal (continuous time-domain)
$s(t)$	Genuine speech signal (continuous time-domain)
$r(n)$	Replay speech signal (discrete time-domain)
$s(n)$	Genuine speech signal (discrete-domain)
$mp(t)$	Impulse response of microphone
$pr(t)$	Impulse response of place of recording
$ms(t)$	Impulse response of multimedia speaker
$pa(t)$	Impulse response of place of attack
F_0	Fundamental frequency
α	Pre-emphasis coefficient
$s_p(n)$	Pre-emphasized signal
T	Total number of frames in utterance
C_{mean}	Cepstral mean
s_t	t^{th} Feature vector
sn_t	Cepstrally mean normalized feature vector
H_0	Hypothesis 0
H_1	Hypothesis 1
λ	Set of GMM parameters
λ_{gen}	GMM of genuine speech
λ_{spoof}	GMM of spoof speech
D	Feature dimension
w_k	Weight of k^{th} component of GMM
μ_k	Mean of k^{th} component of GMM
Σ_k	Covariance matrix of k^{th} component of GMM
$P(s)$	Likelihood function of speech utterance s
LLR	Log-Likelihood Ratio
Sc	Score vector
$d(n)$	Detail part of signal
$m(n)$	Trend part of signal

$e_{up}(n)$	Upper envelope
$e_{low}(n)$	Lower envelope
$m_{env}(n)$	Mean of upper and lower envelope
$d_i(n)$	i^{th} IMF
E	Energy of frame
$G(\omega)$	Group delay function of $s(n)$
$S(\omega)$	Fourier transform of $s(n)$
$Q(\omega)$	Fourier transform of $ns(n)$
$R(\omega)$	Cepstrally smoothed spectra of $S(\omega)$
$G_m(\omega)$	Modified group delay function of $s(n)$
Δ	First order derivative of feature vectors
$\Delta\Delta$	Second order derivative of feature vectors
$p(n)$	Pitch pulse train signal
$h(n)$	Vocal tract impulse response
$N(n)$	Convolution of additional factors in replay speech
$\hat{s}(n)$	Logarithmic cepstrum of $s(n)$
$S(z)$	Z-transform of $s(n)$
$\check{s}(n)$	Spectral root cepstrum of $s(n)$
γ	Exponential factor in spectral root cepstrum
M	Number of filters in filterbank
$MSRCC(q)$	q^{th} MSRCC coefficient
MFM	Mel Frequency Magnitude spectrum
$PSRCC(q)$	q^{th} PSRCC coefficient
MFP	Mel Frequency Phase spectrum
ψ	TEO operator
$\psi(n)$	TEO profile
$\psi_a(n)$	Analytic signal of $\psi(n)$
$\hat{\psi}(n)$	Hilbert transform of $\psi(n)$
$\Psi(\omega)$	Fourier transform of $\psi(n)$
$a_e(n)$	Hilbert envelope
$\phi_\psi(n)$	TEO phase

List of Tables

List of Figures

CHAPTER 1
Introduction

1.1 Introduction

Speech is most powerful and extensively used form of communication among the humans. With the development of technology, one can represent the speech signal into discrete-time, which helps to use speech signal to design many computer applications [7]. The speech signal contains several levels of the information, such as linguistic message, speaker, gender, health, emotion, attribute, acoustic environments, etc. [8]. The automatic speaker recognition aims to identify speaker identity based on the speaker-specific information present in the speech signal. The automatic speaker recognition systems can be further classified into two modes, namely, Speaker Identification (SID) system and Automatic Speaker Verification (ASV) system [9]. In SID the system tries to identify whether the given unknown speaker identity matches with any of the known speaker identities [8]. On the other hand, ASV system verifies the claimed identity of the speaker using the speech samples provided by the speaker [10].

In the past few years, due to significant advancement in speech technology, there is a notable rise in the usage of ASV as a biometric solution for a number of applications, such as telephone banking, attendance, etc. For practical applications, the ASV system needs to be robust against various kind of variabilities, such as a microphone, acoustic noise, transmission channel, intersession, speaker health, speaker aging, etc. Nullifying effect of these components makes the ASV system robust, in addition, it also makes the ASV system susceptible to various kinds of spoofing attacks, which is against the security factor of the practical systems. Hence, for the practical ASV system, it is necessary to develop a solution which nullifies the effect of various kind of variabilities and also make the ASV system secure against malicious spoofing attacks. Fig. 1.1 shows the eight possible attack points in the ASV system. These eight points can be classified as direct attacks (also known as spoofing attacks) and indirect attacks. For indirect spoofing attacks, the attacker must need a access inside the ASV system. The various kinds of spoofing attacks includes impersonation [11], Voice Conversion (VC) [12], Speech Synthesis (SS) [13], twins [14], and replay [15].

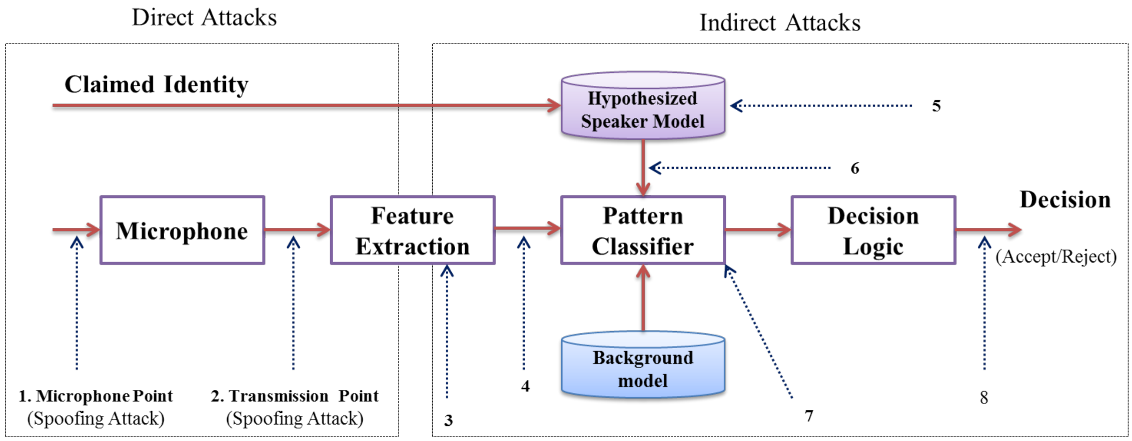


Figure 1.1: Eight possible attack points in the ASV system. After [5].

Impersonation attacks are made by performing human mimicry (human-altered voices), where the attacker tries to imitate the target speaker's identity using human-altered voices without using any computer-aided technology [11]. The accessibility and risk of this kind of attack are low [5]. In VC attacks, the speech samples of a given speaker are manipulated in such a way that, it resembles with the target speakers voice using some computer-aided technologies [12]. VC requires special knowledge of signal processing and a complex algorithm, which makes their accessibility medium. However, advanced VC techniques give very much resembled sound to target speaker, which makes them riskier [5]. SS also referred as Text-to-Speech (TTS), generate the target speaker speech samples for any random text provided using computer-aided technology [13]. The accessibility of speech synthesis is medium, as it requires efficient computers and complex algorithms, however, the risk for this kind of attacks is very high [5]. In twin attacks, the attacker tries to fool the ASV system by providing speech samples of a twin of target speaker [14]. The accessibility of twin attack is low and since there is no detailed study of these attacks, the risk factor is unknown. In the replay attack, an attacker tries to access the speaker's identity by using a pre-recorded speech samples of the target genuine speaker [15]. Replay is the simplest and easily accessible spoofing attack, as it does not require any prior knowledge of specific expertise, a complex algorithm, and special computer knowledge.

1.2 System Architecture

The advanced ASV system may consists of two major components, namely, speaker verification and spoof detection system. Speaker verification assures that whether the input speech signal matches with the claimed identity of the speaker, while the spoof detection system checks for the liveliness of the speech signal, i.e., whether the given input speech signal is natural or spoofed. Fig. 1.2 shows the broader system architecture of Spoof Speech Detection (SSD) system.

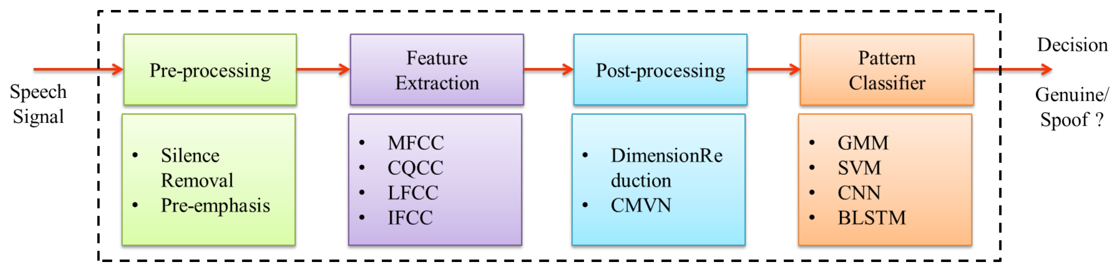


Figure 1.2: System architecture of SSD system.

The details of each block are as follows:

- **Pre-processing:** In this step, the raw speech signal is pre-processed to increase the efficiency of next steps. The various pre-processing techniques used are silence removal, pre-emphasis, etc.

- **Feature Extraction:** The input speech signal is a random process, and cannot be applied directly for classification. Feature extraction maps the input speech signal to new vector space, which gives more sparse and understandable representation of input speech signal. The well-known features used in spoof detection area are Mel Frequency Cepstral Coefficients (MFCC), Constant Q Cepstral Coefficients (CQCC), Instantaneous Frequency Cosine Coefficient (IFCC), etc.

- **Post-processing:** Post-processing techniques are used to map the extracted features to new vector space, where the data can be more discriminative in that space. The post-processing includes channel compensation, feature normalization, dimensionality reduction, etc.

- **Classifier:** Once the speech signal is represented using feature extraction, the next task is to decide whether the input speech signal is natural or spoof. This is done by using various available pattern classifiers. The examples of various pattern classifiers are Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Bidirectional Long Short Term Memory (BLSTM), etc.

1.3 Motivation for Replay Spoof Detection Problem

Among five types of spoofing attacks, replay is the simplest, low effort attack and does not require any sophisticated equipment or specific expertise in the speech signal processing area. Replay attack causes high False Acceptance Rate (FAR) compared to VC and SS attacks (high effort attacks), i.e., in replay attacks, the system is more likely to accept the access by an unauthorized user. The availability of low cost and high quality recording devices, such as mobile phones, makes replay attack most accessible and therefore, great threat to the ASV system. It is quite easy to detect the replay spoofed speech in known conditions compared to unknown conditions. The difficulty of replay spoof detection increases with the high quality intermediate devices, clean recording environment, and playback environment. Because in such cases, the replay speech is very much similar to genuine speech. Hence, the replay speech posses greater risk/threat to the ASV system, thus it is necessary to find out countermeasures for replay spoof detection.

1.4 Applications

The Spoof Speech Detection (SSD) is an integral part of the ASV system. The SSD system has several applications, some of them are listed below:

1. The ASV system is used for numerous applications, such as security, telephone banking, computer and smart phone logins, etc. In all these applications, security is the major issue. Hence, to secure the ASV system against spoofing attacks, the SSD system is necessary.

2. In speaker forensics, it is essential to find out whether the incoming speech signal is from actual suspect or it is an unauthorized access to fool the forensic experts. In such cases, the SSD system can also be used to detect the liveliness of the speech signal.

3. The SSD system can also be used for the evaluation of VC and SS systems. The speech signal generated from VC and SS systems are not recognized by

the SSD system, indicating that the generated signal is more close to natural signal.

4. Depending on the countermeasures proposed, one can investigate the VC and SS algorithms for their drawbacks, and design a new algorithm that will produce more natural speech.

1.5 Contributions in the Thesis

The main objective of this thesis is to develop the replay SSD system using hand-crafted features. In this thesis, three new system-based features are proposed and one source-like feature is explored for replay SSD task.

1.5.1 System-based Features

The human ear process the speech signal in subbands due to the cochlear structure in an ear. Considering this idea for feature extraction, three new system-based features are developed in the thesis.

Empirical Mode Decomposition Cepstral Coefficient (EMDCC) Feature Set

The speech signal is produced from a nonlinear system and it is nonstationary in nature. The Empirical Mode Decomposition (EMD) decomposes the signal into various subbands also known as Intrinsic Mode Functions (IMFs), assuming that the input signal is nonlinear and nonstationary. In addition, EMD also acts as dyadic filterbank. The key idea is that the subbands obtained through the EMD are used to extract the Empirical Mode Decomposition Cepstral Coefficients (EMDCC).

Magnitude-based Spectral Root Cepstral Features

Empirically, it has been observed that the replay speech is more discriminative in the spectral root cepstrum-domain compared to the logarithmic cepstrum-domain. Considering this observation, the Magnitude-based Spectral Root Cepstral Coefficients (MSRCC) are proposed, where the log non-linearity in the MFCC is replaced with the power law nonlinearity.

Phase-based Spectral Root Cepstral Features

The phase information contained in Short-Time Fourier Transform (STFT) is ignored in magnitude-based features, such as MFCC, MSRCC, etc. To use this phase information and capture possible complementary information present in magnitude-based features, the Phase-based Spectral Root Cepstral Coefficients (PSRCC) are proposed. The key difference between the MSRCC and the PSRCC feature sets is the magnitude spectrum is replaced by the unwrapped phase spectrum.

1.5.2 Source-like Feature Set

The excitation source-related information is complementary to speaker-specific information. In this thesis, source-like feature set, namely, Teager Energy Operator (TEO) phase features are explored for replay spoof detection task. Empirically, it has been observed that the TEO phase capture the information which is required for developing high-security ASV system than high user convenience.

1.6 Organization of the Thesis

Figure 1.3 shows the organization of the thesis and details are discussed next:

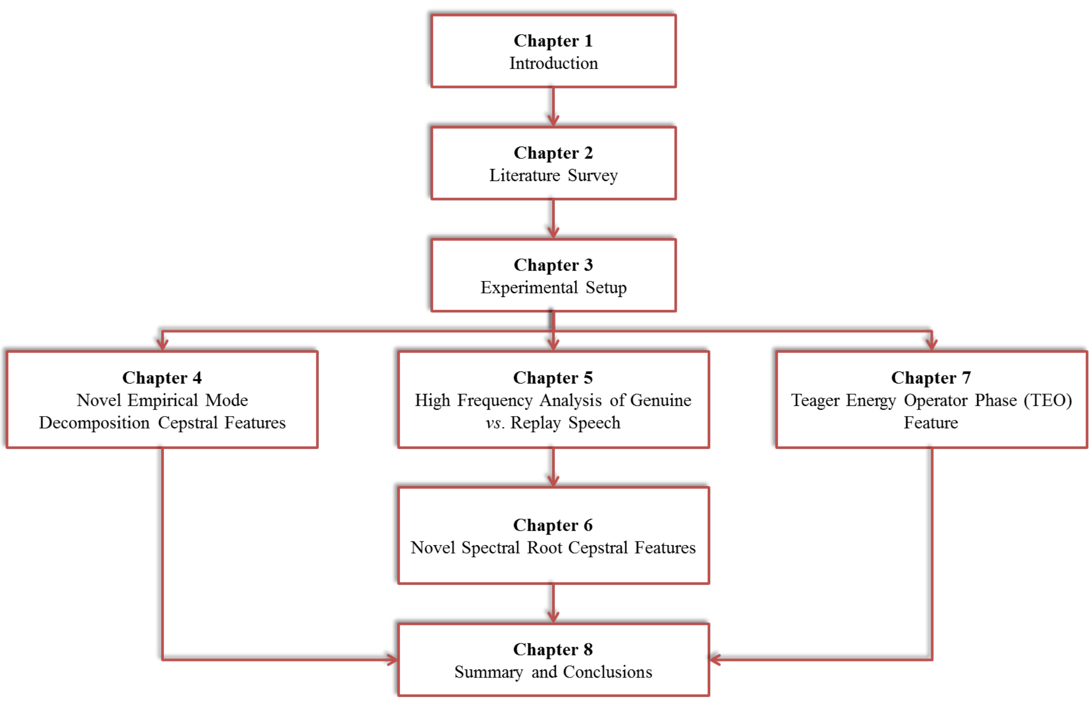


Figure 1.3: Organization of the thesis.

Chapter 2 discusses the literature survey on countermeasures developed for replay spoof detection. The strength and limitations of various replay SSD systems are discussed in this chapter.

Chapter 3 introduces the database, GMM classifier, performance measures, preprocessing and post-processing techniques and data fusion strategy used in the development of replay SSD system.

Chapter 4 presents the newly proposed EMDCC feature set. In this chapter, the brief theory of EMD is described along with some experiments to show that the EMD acts as a dyadic filterbank. The extracted cepstral features are used for developing replay SSD system.

Chapter 5 discuss frequency analysis of genuine *vs.* replay speech using the MFCC and Inverted Mel Frequency Cepstral Coefficients (IMFCC). It illustrates the importance of high frequency for the magnitude spectrum-based features.

Chapter 6 presents the two novel spectral root cepstral feature sets, namely, the MSRCC and the PSRCC. It also discusses the source-filter model of genuine and replay speech along with the difference between logarithmic cepstrum and spectral root cepstrum.

Chapter 7 explores the significance of TEO phase feature for replay spoof detection. It also discusses the effect of TEO phase information on performance evaluation i.e., Detection Error Trade-off (DET) curves.

Chapter 8 concludes and summarizes the work done in this thesis. This chapter also reports the limitations of the current work and suggest possible future research directions.

1.7 Chapter Summary

This chapter introduced the replay spoof detection problem and brief system architecture of replay SSD system is presented. In this chapter, the motivation for replay spoof detection problem, applications of SSD system, contributions in the thesis and organization of the thesis is discussed. In the next chapter, details of literature survey of replay spoof problem along with strengths and limitations of various approaches is presented.

CHAPTER 2
Literature Survey

This chapter presents the brief literature survey on countermeasures developed for replay spoof detection. The mathematical model simplifies the analysis and hence, in this chapter, the mathematical model of replay speech is discussed. The comparison of some key landmark replay SSD systems is done with respect to features, classifier, results, strength and limitation of the systems. In the last section, the research issues in replay spoof detection task are presented.

2.1 Mathematical Model of Replay Speech Signal

The replay speech is convolution of the genuine speech and impulse response of microphone, impulse response of multimedia speaker, impulse response of place of recording, impulse response of place of attack [16]. Hence, the replay speech can be written as:

$$r(t) = s(t) * mp(t) * pr(t) * ms(t) * pa(t), \tag{2.1}$$

where $mp(t), ms(t), pr(t)$ and $pa(t)$ represents impulse response of microphone, impulse response of multimedia speaker, impulse response of place of recording and impulse response of place of attack, respectively.

2.2 Literature Review

One of the initial study in replay SSD for the text-dependent system was reported in [17]. To verify the input speech, a choice made based on a set of N similarity scores is used. The channel noise from recording device was used to detect the replay speech [18]. The replay speech obtained through far-field recording using land line and GSM telephone channel using modulation index and spectral ratio was studied in [19] for text-dependent ASV system and in [20] for text-independent ASV system. For text-dependent ASV system, the spectral bitmaps are used to determine whether the input speech is natural or replayed [15]. In [21], similar technique of average spectral bitmaps is used for text-independent ASV system.

ASV spoof 2017 challenge is organized to develop the countermeasures of the replay attack detection on highly heterogeneous recording and replayed conditions. Baseline system uses GMM as back-end classifier with CQCC feature set that is based on the perceptually-motivated time-frequency transform [1]. Study reported in [22] investigated different spectral features, such as CQCC, MFCC, etc. and their feature-level fusion. The cross-database experiments with the BTAS 2016 ASVspoof development set are also reported [22]. A new feature extraction approach, namely, Variable length Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients (VESA-IFCC) is proposed to exploit the usefulness of Instantaneous Frequency (IF) in subband energy via Energy Separation Algorithm (ESA) [23]. The study in [24] has explored different feature set as Glottal Closure Instants (GCIs), epoch strength and the peak-to-sidelobe ratio of Hilbert envelope of the Linear Prediction (LP) residual. The data augmentation technique along with deep Residual Network (ResNet) is used in [25]. The same study also shows experiment using Deep Neural Network (DNN), BLSTM neural network as classifiers with CQCC features.

As replay spoofing attack affects on the high frequency spectrum, analysis using inverse-MFCC, Linear Prediction Cepstral Coefficients (LPCC), and LPCC-residual (LPCCres) features is performed in high frequency regions [26]. In [27], feature selection methods are applied on mean and variance of CQCC features with SVM as a classifier. DNN architectures, such as Light CNN , CNN + Recurrent Neural Network (RNN) are used along with *i*-vector+SVM, Constant Q Transform (CQT), and FFT as feature representation [28]. The ensemble classifier set using multiple GMM, GMM mean supervector-Gradient Boosting Decision Tree (GSV-GBDT) and GSV-Random Forest (GSV-RF) classifiers is proposed in [29]. An F-ratio probing tool is used to analyze the three variability factors, i.e., speaker identity, speech content and playback and recording device [30]. In these set of experiments, it is observed that replay device contributes to *overfitting* risk. The fusion of high level features using DNN with the CQCC and High Frequency Cepstral Coefficients (HFCC) is investigated in [31]. Model fusion using GMM, DNN, and ResNet is performed in [32]. It is found that multi-channel information in replayed speech is found in low SNR regions. The analysis is performed using Single Frequency Filtering (SFF) [33].

Table 2.1: Summary of the countermeasures proposed for the ASV Spoof 2017 challenge

Study	Features	Classifier	EER (in %)	Strengths	Limitations
Kinnunen *et. al.* [1]	CQCC	GMM	30.60 %	Use of CQT	Higher dimension of feature vector compared to MFCC
Patil *et. al.* [23]	VESA-IFCC	GMM	14.06 %	ESA to capture possible changes in spectral envelope	The information in amplitude envelope is completely neglected
Jelil *et. al.* [24]	EF, PSRMS, IFCC, CQCC, MFCC	GMM	13.95	Source-based features to understand dissimilarities	Difficult to find exact position of GCIs
Cai *et. al.* [25]	CQCC, representation learning	GMM, FDNN, BLSTM	16.39	Artificial data augmentation strategy, end-to-end representation learning	Data augmentation is done manually, STFT is used as input
Witkowski *et. al.* [26]	IMFCC, LPCC, LPCCres	GMM	17.31	The strongest generalization to unseen data	Lower frequency information completely ignored
Lavrentyeva *et. al.* [28]	Magnitude spectrum through FFT and CQT	SVM, LCNN, CNN, RNN	6.73	Combination of CNN, RNN, and CQT spectrogram	Perform better for phrase-dependent than phrase-independent
Nagarsheth *et. al.* [31]	CQCC, HFCC	GMM, SVM, DNN	11.5	DNN is trained to maximize the channel variability	Uses low-level features

Study	Features	Classifier	EER (in %)	Strengths	Limitations
Alluri *et. al.*	SFFCC	GMM, BLSTM	17.82	High spectro-temporal resolution to extract robust features	No consistency in the feature combination across dimensions
Chen et. al.	CQCC, MFCC	DNN, ResNet	13.30	ResNet and model fusion	Feature diversity does not perform better
Ji. *et. al.*	CQCC	GMM, GMM-UBM, GSV-SVM, GSV-SVM-NAP, ivector-GPLDA, GSV-RF, GSV-GBDT, ensemble learning	12.4	Effective post-processing method, multiple classifiers	Data *Overfitting* problem in training

2.3 Research Issues in Replay Spoof Detection

Some of the research issues in replay spoof detection problems are as follows:

- The quality of recording/ playback device and noise level in recording/ playback environment decides the quality of replay speech and hence, the difficulty of replay spoof detection. The replay speech recorded with a high-quality recording and playback device in clean recording environment is very difficult to detect as it is very much similar to the genuine speech.

- The performance of countermeasures developed for replay spoof detection problem is the measure for certain database. The database contains the limited number of recording, playback devices, limited numbers of recording and playback environment. Hence, the performance of the systems under the unseen conditions is unknown.

- The performance of spoof detection system is generally considered in FAR, however, the features should provide higher user convenience, i.e., lower False Rejection Rate (FRR) along with high security.
- The conventional features, such as MFCC, LPCC are designed for speaker recognition and hence, there is need to develop the features especially for spoof detection problem.

2.4 Chapter Summary

In this chapter, the mathematical model of replay speech with respect to the genuine speech is discussed. The brief literature survey on countermeasures developed for replay spoof detection is presented. Furthermore, we compared some key landmark replay SSD systems with respect to features, classifier, results, strength and limitation of the systems. At last, the research issues in replay SSD task is discussed. In the next chapter, the common experimental setup, such as a database, classifier, pre-processing and post-processing techniques, performance measures and data fusion strategy is discussed.

CHAPTER 3

Details of Spoof Speech Detection (SSD) System Framework

This chapter discusses the database, classifier, pre-processing, post-processing techniques, performance measures, and data fusion strategy used in the development of replay SSD systems. This chapter gives the statistics of the ASV spoof 2017 challenge database. In this chapter, the GMM classifier for binary hypothesis problem is discussed. In addition, it also gives an idea about pre-emphasis (pre-processing technique) and CMN (post-processing technique). Furthermore, the DET curve (performance measure) is discussed in detail and finally, the data fusion strategy is briefly explained.

3.1 Database Used

All the experiments are conducted on statistically meaningful and publicly available ASV spoof 2017 challenge database [1]. The full database contains three subsets, namely, training set, development set, and evaluation set.

Table 3.1: Summary of ASV spoof 2017 challenge database. After [1]

Subset	**No. of Speakers**	**No. of Replay Sessions**	**No. of Replay Config.**	**No. of Utterances**	
				Natural	**Spoof**
Training	10	6	3	1508	1508
Development	8	10	10	760	950
Evaluation	24	161	110	1298	12008

The database is primarily based on the recent text-dependent RedDots corpus [34] and its replayed version [35]. It contains 10 common phrases, and the recording was conducted with different playback and recording devices. The sampling rate of each utterance is 16 kHz and the sample precision is 16-bits. The details of the database are given in [1]. Table 3.1 summarizes brief details of this database.

3.2 Pre-processing and Post-processing Techniques

The real-time speech is a raw signal and may not be efficient for the feature extraction and classification. To increase the efficiency and to convert this raw signal into a refined signal, pre-processing and post-processing techniques are used.

3.2.1 Pre-processing Techniques

Pre-processing techniques are used before feature extraction in order to increase the efficiency of feature extraction process. The various pre-processing techniques used are silence removal, pre-emphasis, etc. In some of the experiments, pre-emphasis technique is used as the pre-processing technique.

Pre-emphasis

The replay speech contains more discrimination cues in high frequency range [26]. In the pre-emphasis, the high-frequency signal components are emphasized. The transfer function of pre-emphasis filter is given as:

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

where α is pre-emphasis coefficient and it decides the cutoff frequency of the filter.

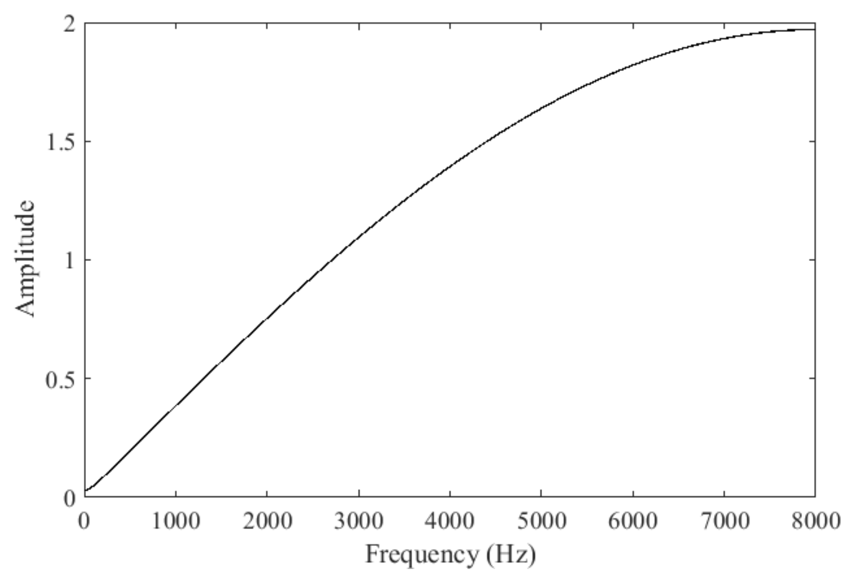


Figure 3.1: Magnitude response of the pre-emphasis filter for $\alpha = 0.97$.

Let the $s(n)$ is a speech signal, then the pre-emphasized signal $s_p(n)$ is given as:

$$s_p(n) = s(n) - \alpha s(n-1). \tag{3.2}$$

For most of the speech applications α value is taken as 0.97. Figure 3.1 shows the magnitude response of pre-emphasis filter for $\alpha = 0.97$.

3.2.2 Post-processing Techniques

Post-processing techniques are used after feature extraction and before classification stage. Post-processing techniques map the extracted features to new vector space, where the data is more discriminable. The post-processing includes channel compensation, feature normalization, dimensionality reduction, etc. In this thesis, some of the experiments use CMN technique (feature normalization technique).

Cepstral Mean Normalization (CMN)

During speech recordings, the characteristics of a transmission channel may vary from one session to another. To minimize the effect of these variations CMN is performed. In CMN, the mean of cepstral coefficient over an utterance is subtracted from each frame. Let $U = \{s_1, s_2, ..., s_T\}$ represents T feature vectors of utterance U, then the cepstral mean is given by:

$$C_{mean} = \frac{1}{T}\sum_{t=1}^{T} s_t. \tag{3.3}$$

The normalized feature vectors, $U_n = \{sn_1, sn_2, ..., sn_T\}$ are computed as,

$$sn_t = s_t - C_{mean}, \qquad t = 1, 2, ..., T. \tag{3.4}$$

3.3 Pattern Classifier

Given a speech signal S, replay spoof detection task is to decide, whether S is a genuine speech or replay speech. The replay spoof detection can be stated as binary hypothesis testing between H_0 and H_1, where hypothesis H_0 and H_1 are given as follows,

H_0: S is a genuine speech (represented as λ_{gen})
H_1: S is a replay speech (represented as λ_{spoof})

3.3.1 Gaussian Mixture Model (GMM)

A GMM is a weighted sum of Gaussian distributions. The D-dimensional feature vector, s and K component Gaussian mixture density are used for the likelihood function and is defined as [36]:

$$P(s) = \sum_{k=1}^{K} w_k P_k \left(s|\mu_k, \Sigma_k\right), \tag{3.5}$$

where $P_k(s)$ is Gaussian distribution, parameterized by a $D \times 1$ mean vector (μ_k), and a $D \times D$ covariance matrix (Σ_k).

$$\therefore P(s|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(s-\mu_k)^t \Sigma_k^{-1}(s-\mu_k)}. \tag{3.6}$$

The parameters of the GMM can be trained using Expectation Maximization (EM) algorithm. The details of EM algorithm are given in Appendix A.

3.3.2 Log-Likelihood Detector

For the feature vectors, $S = \{s_1, ..., s_T\}$ of speech utterance, S having T number of frames. Assuming independence of feature vectors, the log-likelihood is computed as:

$$\log p(S|\lambda) = \frac{1}{T} \log p(s_1, ..., s_T|\lambda), \tag{3.7}$$

$$\log p(S|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(s_t|\lambda), \tag{3.8}$$

where λ denotes the parameters of GMM, i.e., $\lambda = \{w_k, \mu_k, \Sigma_k\}$ for $k = 1, 2, ..., K$. Once the log-likelihood is computed of each hypothesis for a given utterance, the Log-Likelihood Ratio (LLR) is computed as follows:

$$\Lambda(S) = \log p(S|H_0) - \log p(S|H_1), \tag{3.9}$$

$$\Lambda(S) = \log p(S|\lambda_{gen}) - \log p(S|\lambda_{spoof}), \tag{3.10}$$

where λ_{gen} is GMM of genuine speech and λ_{spoof} is GMM of replay spoof speech. The log-likelihood detector decides the class of S comparing this LLR score with pre-determined threshold value. If the LLR score is greater than threshold S is declared as genuine speech else S is declared as replay spoof speech.

3.4 Performance Measure

3.4.1 Detection Error Trade-off (DET) Curve

Replay SSD task is a binary hypothesis testing problem. It involves a trade-off between two types of error, namely, type-I and type-II. Type-I errors happen when system rejects the true hypothesis. Type-I errors are also known as false positive, false rejection or missed detection. Type-II errors happen when system accept a hypothesis which is actually false. Type-II errors are also known as false negatives, false acceptance, or false alarm. Systems that involves the trade-off between two types of error are not able to describe their performance/strength using a single operating point. Rather, such system has several operating points and can be characterized by a performance curve [6]. Traditionally, two types of performance curves are used to describe the system performance, namely, Receiver Operating Characteristics (ROC) curve and Detection Error Trade-off (DET) curves.

The ROC curve is usually used to measure the performance of detection systems. ROC uses linear scales, in which correct detection rate (Y-axis) is plotted as a function of false alarm rate (X-axis). ROCs are the nonlinear curve, which makes them difficult to analyze. The upper left point of the plot is the optimal point in ROC and the curve of the well performing systems lead to bundle together in upper left point [6].

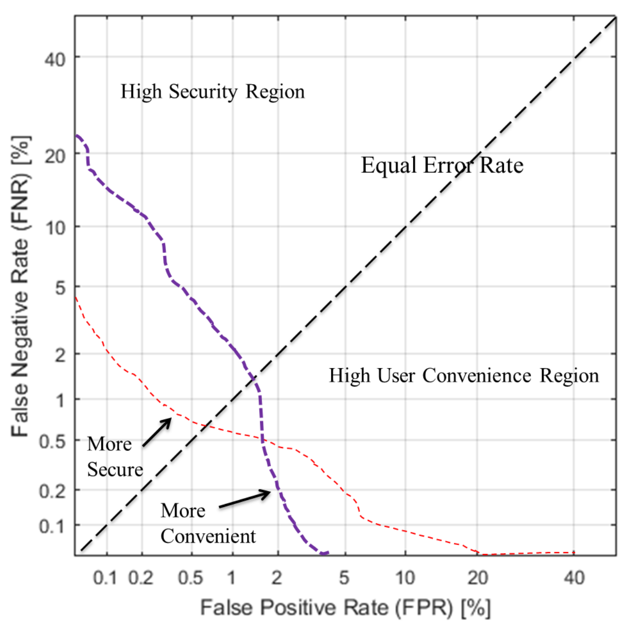


Figure 3.2: Examples of DET curve and important regions in DET curve. After [6].

The DET curve is derived from the ROC curve. The nonlinearity in ROC curve makes them difficult to analyze. The linear curves are easy to analyze hence, in the DET curve both the error rates are plotted on horizontal and vertical-axis, giving uniform weight to type-I and type-II errors. The DET curves use a scale for both the axes which spreads out the curve and better discriminate the different well performing systems, producing an approximately linear plot. The DET curve uses normal deviate scale or simple logarithmic scale [6]. Figure 3.2 shows the example of DET curves and important regions.

3.4.2 Equal Error Rate (EER)

Equal Error Rate (EER) is a special point on the DET curve. EER is defined as an operating point at which false acceptance rate is equal to false rejection rate. Lower EER indicates that the system is performing well and accurate [6].

3.5 Data Fusion Strategy

In this thesis, score-level data fusion strategy is used. The advantage of score-level fusion is that it gives the idea about individual feature contribution towards system performance without increasing feature dimension. Consider Sc_1 and Sc_2 are the two score files obtained from two different feature sets, then the score-level fusion is given by:

$$Sc_{new} = \alpha_f Sc_1 + (1 - \alpha_f) Sc_2, \tag{3.11}$$

where α_f is fusion factor and Sc_{new} is new score obtained. If $\alpha_f > 0.5$, then first feature set contribute more towards performance of the system otherwise second feature set contributes more.

3.6 Chapter Summary

In this chapter, the detailed experimental framework of SSD system used to developed replay SSD system is discussed. The chapter discussed the database, GMM classifier, pre-processing and post-processing techniques, performance measures and data fusion strategy. In the next chapter, novel empirical mode decomposition cepstral feature set is discussed for replay SSD task.

CHAPTER 4

Novel Empirical Mode Decomposition Cepstral Features

Several conventional features, such as MFCC, LFCC, IFCC, etc. use filterbank structure to obtain the subband filtered signal for further processing. This filterbank is fixed for all the utterances in feature extraction process. In addition, these feature extraction methods assume that the input speech signal is produced by a linear systen and it is a stationary signal [37]. Some of the studies used the EMD for speech applications, such as formant frequency estimation [38], noise reduction [39]. The EMD was first used for nonlinear and nonstationary time-series analysis [40]. The EMD decomposes the signal into special IMFs, assuming that the input signal is non-linear and non-stationary and all the events are handled as they arise. Flandrin *et. al.* reported that the EMD acts as a dyadic filterbank and decomposes input signal similar to wavelet-like decomposition [41]. In this chapter, a new feature extraction approach is discussed in which the filterbank is replaced with EMD as it can consider the nonlinear and non-stationary nature of the speech signal. This new proposed feature set in this thesis is referred to as EMDCC. The experimental results are compared with MFCC [42] and CQCC [43] results. Furthermore, EMDCC feature set is fused with Linear Frequency Modified Group Delay Cepstral Coefficients (LFMGDCC) at score-level to capture the possible complementary information.

4.1 Empirical Mode Decomposition (EMD)

The EMD deals with the signals that are nonlinear and nonstationary and the events are handled as the signal occur [44]. EMD decomposes a signal into zero mean Amplitude Modulation and Frequency Modulation (AM-FM) waveforms. These AM-FM waveforms are known as IMFs, and aims to represent underlying intra-wave modulated components in the signal. The *locally* zero mean condition is assured by maintaining the mean of the lower and upper envelope *(obtained by interpolating minimas and maximas)* of an IMF equal to zero [40, 45]. For a given signal $s(n)$, it can be decomposed into two parts, one is high-frequency (local) part also known as *detail*, $d(n)$, and the other part is low-frequency (local) part known as *trend*, $m(n)$ [2, 41]:

$$s(n) = d(n) + m(n). \tag{4.1}$$

To obtain the further *detail* and *trend* signals, the procedure of decomposing signal into *detail* and *trend* is iteratively applied on residual term considering as a new signal. For a given signal, $s(n)$, the IMFs can be found as follows [40]:

1. Identify all the relative maxima and minima of the signal.

2. Interpolate maxima (and minima) to get upper envelope $e_{up}(n)$ (and lower envelope $e_{low}(n)$).

3. The mean of upper and lower envelope is computed as [2]:

$$m_{env}(n) = \frac{e_{up}(n) + e_{low}(n)}{2}. \tag{4.2}$$

4. To get rid of overriding wave, the mean signal $m_{env}(n)$ is subtracted from the original signal $s(n)$ [2]:

$$x_1(n) = s(n) - m_{env}(n). \tag{4.3}$$

   Due to small humps present in speech signal which are not identified as maxima or minima in the first step, $x_1(n)$ cannot be said as IMF. Hence, to get the actual IMF, the process continued iteratively, considering $x_1(n)$ as new signal, sifting process is continued to obtain $x_{11}(n)$. Following series is called as $x_{1j}(n)$. The procedure is repeated till the following condition is met by the residue [2]:

$$\sum_{n=0}^{N} \left[\frac{|x_{1(j-1)}(n) - x_{1j}(n)|^2}{(x_{1(j-1)}(n))^2} \right] \leq 0.2. \tag{4.4}$$

   Assuming $d_1(n) = h_{1J}$, satisfies condition in Eq. (4.4), $d_1(n)$ is called as the first IMF. The Left-Hand Side (LHS) of Eq. ([**?**]) indicates the standard deviation term. Lesser standard deviation indicates that the signals are more identical to each other. Here, the standard deviation between two consecutive residues is computed, hence when the residue term does not change much more, the iterative process is stopped. There is no mathematical justification for threshold value 0.2. However, experimental results show that the threshold value 0.2 is sufficient.

5. Subtract the IMF $d_1(n)$ from the actual signal $s(n)$ [2]:

$$m_1(n) = s(n) - d_1(n). \tag{4.5}$$

6. Considering this, $m_1(n)$ as a new signal, the entire process (step 1- step 5) is repeated to obtain other IMFs. For stopping either of the following two conditions should be satisfied.

   - **Condition A:** The signal does not contain any maxima and minima i.e., signal becomes void or monotonic, or
   - **Condition B:** The signal level becomes negligible such that it can be neglected for further processing.

Finally, for a given signal $s(n)$, having I number of IMFs the EMD can be represented as [2]:

$$s(n) = m_I(n) + \sum_{i=1}^{I-1} d_i(n), \tag{4.6}$$

where $m_I(n)$ represents residual term and $d_i(n)$, $(i = 1, 2, ..., I-1)$ represents zero-mean AM-FM waveforms (IMFs/modes).

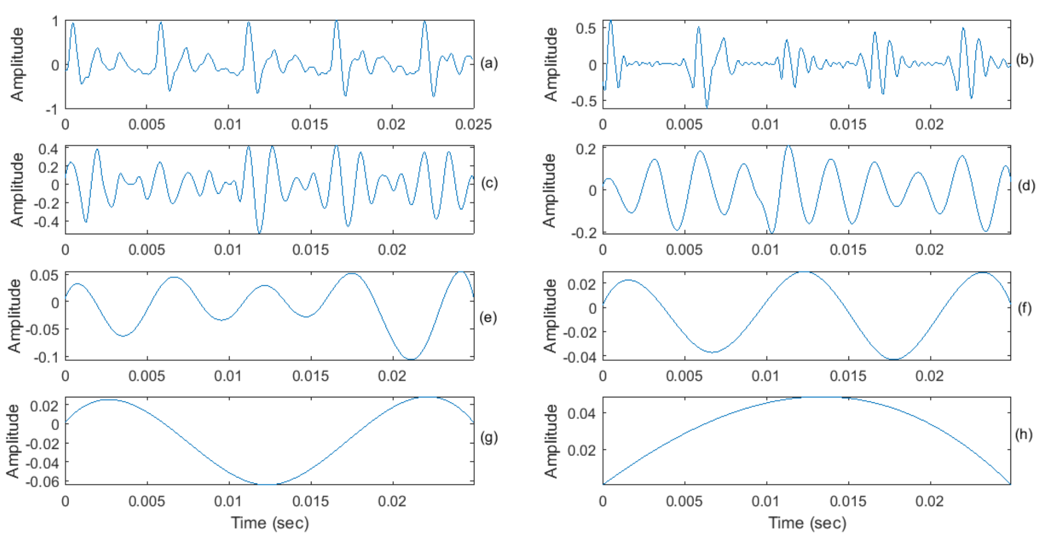


Figure 4.1: (a) Voiced frame of speech signal, (b-g) IMF: 1-6 and (h) IMF: 7 or trend. After [2].

Figure 4.1 shows the IMFs obtained for the voiced frame of the speech signal. The first IMF obtained is high frequency signal which indicates that the filter for IMF: 1 (or mode: 1) is high pass in nature, while the other filters associated with another IMFs or modes are bandpass in nature [41]. The narrowband condition for decomposed signals can be assured by maintaining number of zero-crossings of components to be either same or differ by at most one with number of minima and maxima.

The EMD behaves as a dyadic filterbank and decomposes signal similar to the wavelet-like decomposition for fractional Gaussian noise [41] and for white noise [46]. From the process of EMD, each IMF is zero-mean AM-FM waveform whose number of zero-crossings is equal or differ by at most one with its number of extremas. The number of zero-crossings can be roughly interpreted as mean frequency of IMF. Figure 4.2 shows the comparison of zero-crossings (mean frequencies) of each IMF and center frequencies of subbands for log_2, Mel, Bark, and linear scale. Empirical results shows that the number of zero crossing resemble to log_2 scale and show equivalent structure of filterbank with log scale [41].

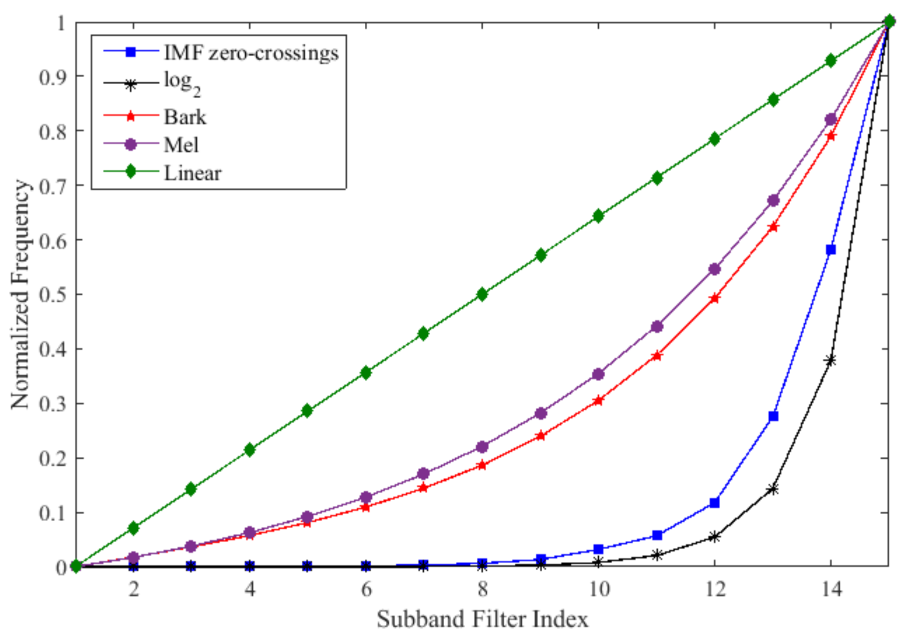


Figure 4.2: Comparison of the zero-crossings (mean frequencies) of IMF and center frequencies of subbands for log_2, Mel, Bark, and linear scale.

Figure 4.3 shows the spectrogram of first 11 IMFs obtained by EMD decomposition and Figure 4.4 shows the spectrogram of first 11 subbands obtained by passing speech signal through mel triangular filterbank of 11 subband filters.

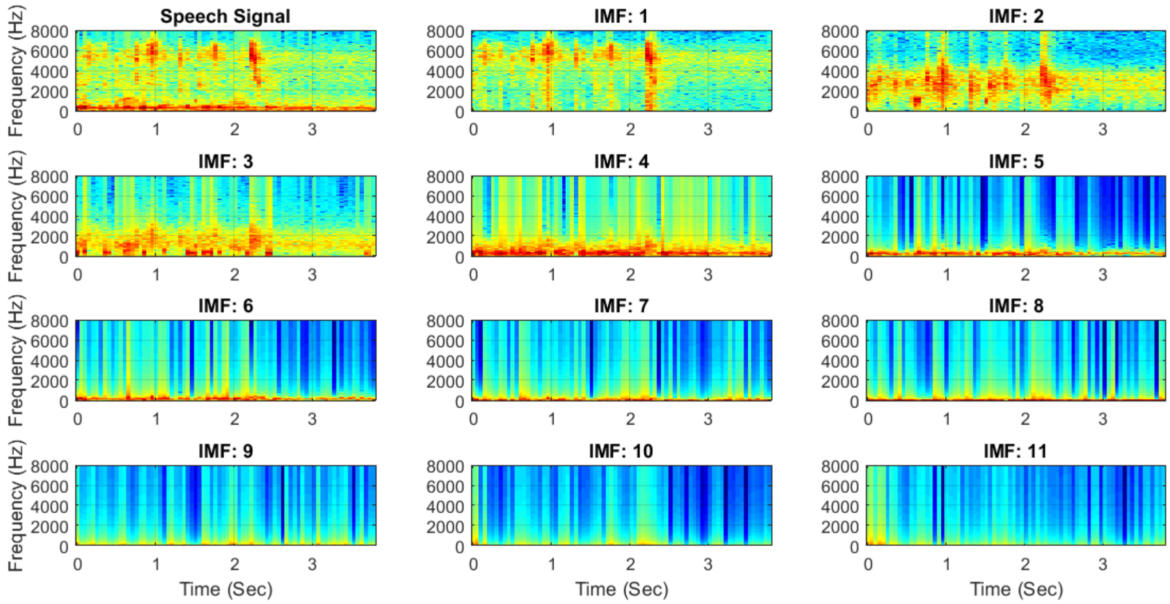


Figure 4.3: Spectrogram of first 11 IMFs obtained through EMD decomposition. After [2].

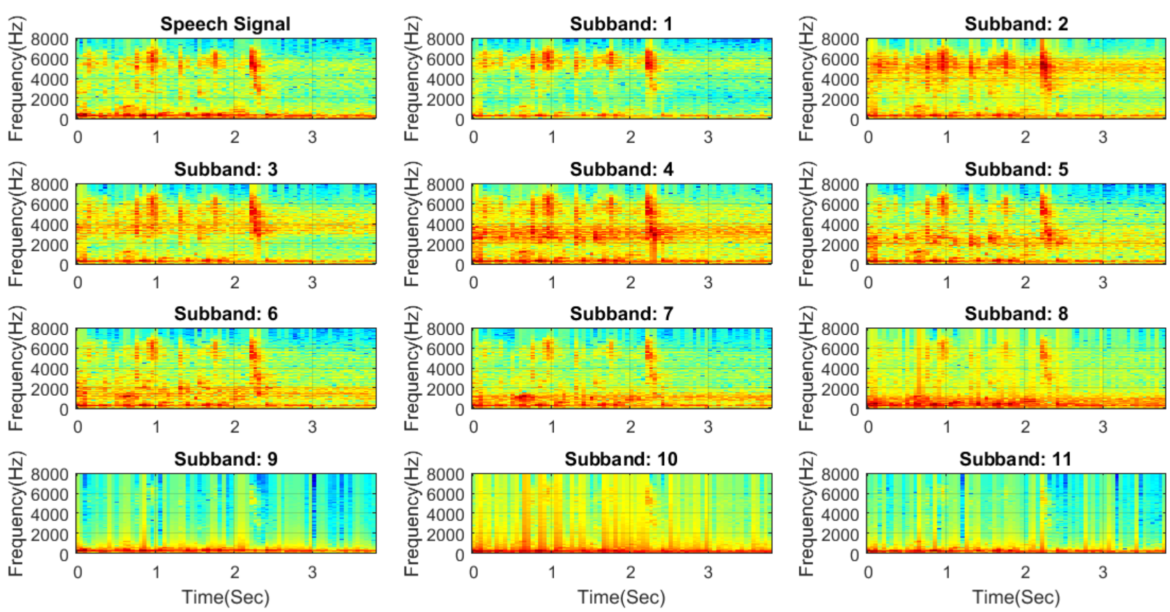


Figure 4.4: Spectrogram of first 11 subbands obtained through triangular filterbank. After [2].

From Figure 4.3 and Figure 4.4, it is observed that the IMF spectrogram is very much similar to the spectrogram of subband obtained through filterbank. However, EMD gives the very good resolution at the lower frequencies and poor resolution at the higher frequencies compared to the filterbank *(due to log_2 scale)*. It is also observed that EMD behaves as a very sharp filter blocking for most of the frequencies outside the range of IMF frequencies compared to the filterbank, which passes other frequencies with less attenuation.

4.2 Proposed Feature Extraction

Figure 4.5 shows the functional block diagram to extract the proposed EMDCC feature set. The input speech signal is first decomposed into I zero-mean AM-FM waveforms (IMFs). Each IMF is then segmented with 20 ms duration with 50 % overlap. Furthermore, the energy in each frame $x(n)$ is computed for every IMF as follows [2]:

$$E = \sum_{m=0}^{M-1} |x(m)|^2, \tag{4.7}$$

where M is the length of frame (320 samples) and E represents energy in frame $x(n)$. To approximate the nonlinear relation between auditory nerve firing rate and signal intensity, log function is applied on the computed energies. This non-linearity provides remarkable robustness by suppressing small signal variability and also mimics the human perception of loudness [45, 47].

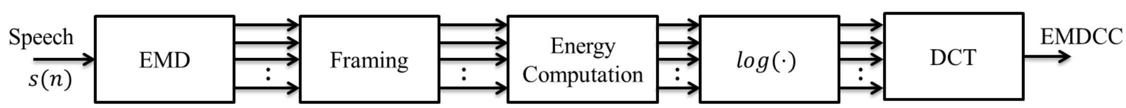


Figure 4.5: Functional block diagram to extract the proposed EMDCC feature set. After [2].

The Discrete Cosine Transform (DCT) is used to reduce the redundancy amongst feature (i.e., decorrelation) and to obtain the cepstral coefficients. First few cepstral coefficients are retained and in order to capture the dynamic nature of the speech signal, they are appended with their first and second-order derivatives to obtain higher-dimensional feature vector.

4.3 Linear Frequency Modified Group Delay Cepstral Coefficients (LFMGDCC)

The phase spectrum need complex unwrapping algorithm before processing. Hence, the use of phase spectrum in speech applications has been ignored for several years. In this replay SSD system, to capture the phase information, LFMGDCC feature set is used with EMDCC feature set. The Group Delay (GD) function, which is defined as negative frequency derivative of phase function, has same properties as the phase function [48]. The group delay indicates the time delay of the amplitude envelopes of the various sinusoidal components of a signal through a system, and is a function of frequency for each component. Speech signal is an output of a stable system and hence, only zeros (due to noise or analysis window) are very close to the unit circle. These zeros causes the GD function to be spiky in nature. The Modified Group Delay (MGD) function, proposed in [49] suppresses these zeros. The GD function $G(\omega)$ for speech frame $s(n)$ can be computed as [2, 50]:

$$G(\omega) = \frac{S_r(\omega)Q_r(\omega) + S_i(\omega)Q_i(\omega)}{|S(\omega)|^2}, \tag{4.8}$$

where $S(\omega)$ and $Q(\omega)$ represents Fourier transform of $s(n)$ and $ns(n)$, respectively (suffix r and i indicates real and imaginary part respectively). To suppress the zeros close to the unit circle, the denominator term in Eq. (4.8) is replaced by cepstrally smoothed spectra ($R(\omega)$) of $S(\omega)$ [2], i.e.,

$$G(\omega) = \frac{S_r(\omega)Q_r(\omega) + S_i(\omega)Q_i(\omega)}{|R(\omega)|^{2\rho}}, \tag{4.9}$$

$$G_m(\omega) = \frac{G(\omega)}{|G(\omega)|}(|G(\omega)|^{\gamma}), \tag{4.10}$$

where $G_m(\omega)$ represents MGD function. The parameters ρ and γ decides the reduction level in amplitude of spikes and restores the dynamic range in GD function and need to be tuned as per application. To use this MGD function for replay spoof detection, the cepstral coefficients are computed similar to the Mel Frequency Modified Group Delay Cepstral Coefficients (MFMGDCC) [51] except that the Mel scale is replaced with linear scale and refer this feature set as LFMGDCC. The extraction of LFMGDCC start with computing MGD values and passing them through linear triangular filterbank. The energy in each filter of the filterbank is computed, and to decorrelate the signal DCT is applied [52]. First, few coefficients are retained and are appended with Δ and $\Delta\Delta$ coefficients to capture dynamic information.

4.4 Feature Parameterization

4.4.1 EMDCC

The cubic spline interpolation is used to interpolate maxima and minima to obtain upper and lower envelope, respectively. The first 10 IMFs obtained from EMD are used for the feature extraction. The 10 IMFs are segmented into short frame of duration 20 *ms* with 50 % overlap. The 10 static coefficients are appended with Δ and $\Delta\Delta$ coefficients, resulting in *30*-dimensional (D) feature vector. The GMM classifier with 512 Gaussian components is used.

4.4.2 LFMGDCC

The parameter ρ and γ are set to *0.4* and *0.1*, respectively, for computing MGD function. The 13-D static LFMGDCCs along with Δ and $\Delta\Delta$ coefficients are used to get 39-D feature vector. The feature coefficients were extracted using 40 linearly scaled triangular filters along with the Hamming window of 25 *ms* duration and 10 *ms* shift. The GMM classifier with 512 Gaussian components is used.

4.5 Experimental Results

The results of the proposed feature set along with other feature sets are shown in Table 4.1, where '+' indicates the score-level fusion. The proposed feature set results are compared with CQCC (baseline system) and MFCC feature set.

Table 4.1: Results on development and evaluation dataset. After [2]

Feature Set	EER (%) Development Set	EER (%) Evaluation Set
CQCC *(Baseline)*	**12.11**	29.18
MFCC	12.21	31.3
LFMGDCC	19.26	22.91
EMDCC	28.48	28.06
EMDCC+LFMGDCC	12.42	**18.36**

'+' indicates score-level fusion

The CQCC-GMM system is a baseline system provided by the organizers of the ASV Spoof 2017 Challenge, having an EER of 29.18 % on evaluation set. The

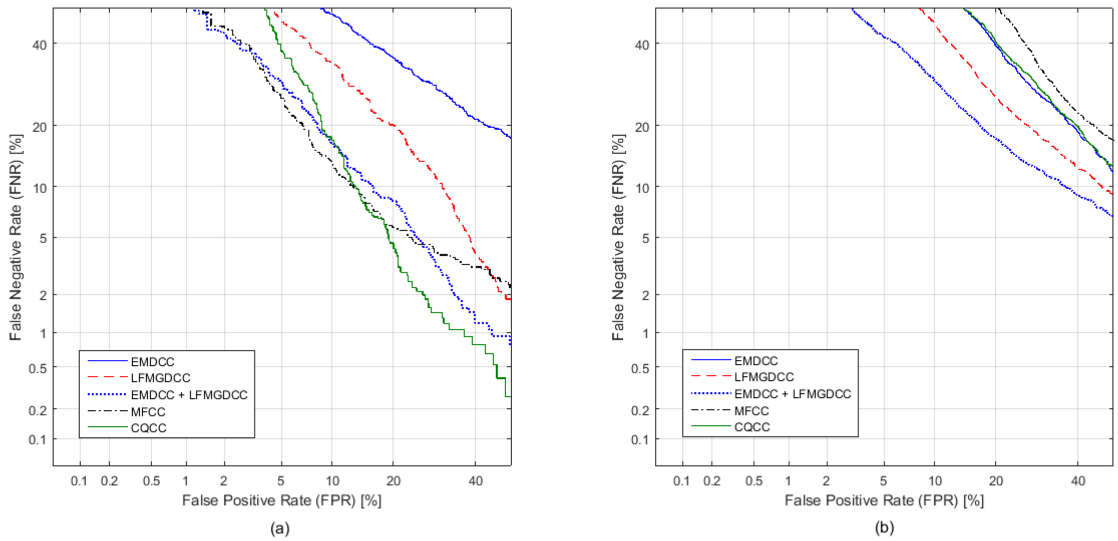


Figure 4.6: DET curves for (a) development, and (b) evaluation set. After [2].

MFCC-GMM and LFMGDCC-GMM systems gives an EER of 31.3 % and 22.91 %, respectively, on evaluation set. The EMDCC gives result of 28.06 % compared to the MFCC system on evaluation set giving improvement of 3.24 %. Furthermore, EMDCC feature set is fused with phase-based LFMGDCC feature at score-level giving reduced EER of 12.42 % and 18.36 % on development and evaluation set, respectively, indicating that the proposed feature set captures complementary information. Figure 4.6 shows the DET curves for the development and evaluation set of the ASV spoof 2017 challenge database. From DET plots, it can be observed that score-level fusion of EMDCC and LFMGDCC feature sets gives better performance at all the operating points of the DET curve. Some experiments also performed with the power law nonlinearity, where the exponent value of $-1/30$ gives best results (i.e., least % EER). Using power law nonlinearity, the EER of 28.99 % and 27.87 % is obtained on development and evaluation set of ASV spoof 2017 challenge database, respectively. However, it is observed that log nonlinearity captures better complementary information compared to the power law nonlinearity.

4.6 Chapter Summary

In this chapter, a novel EMDCC feature set for replay spoof detection task is discussed. The EMD decomposes a signal into IMFs assuming that the signal is nonlinear and nonstationary and acts as a dyadic filterbank. The proposed feature set uses the EMD to decompose speech signal into subbands and uses log nonlinearity to approximate nonlinear relation between auditory nerve firing rate and signal intensity. Furthermore, this feature set is fused with phase-based feature set LFMGDCC at score-level, to capture the possible complementary information. The individual EMDCC system performance for power law nonlinearity is better, however, log nonlinearity is found to capture better complementary information. The results of the final fused system are compared with the baseline CQCC and MFCC systems. Although EMDCC feature set gives comparative EER with the baseline systems, the number of IMFs obtained for signal is decided by the length of the signal. Hence, it is necessary to have input speech signal of fixed duration say 2/3 sec. If the signal is greater than this then the 2/3 sec signal can be cropped. For the database, we can use the fixed length as minimum signal duration present in database. For practical applications we need to decide some fixed duration. There is no mathematical expression to find out fixed duration. In the next chapter, the frequency range analysis for the genuine and replay speech signal is carried out.

CHAPTER 5

High Frequency Analysis of Genuine *vs.* Replay Speech

To design the replay SSD system, it is necessary to know the distinguishable factors between genuine and replay speech. From the mathematical model, one can say that the replay speech is a convolution of the genuine speech signal with additional noise component due to replay mechanism. In most of the cases, this replay noise is high frequency signal and the speech signal contains more information at the lower frequencies compared to the higher frequencies. In this chapter, empirically it is proved that the replay speech signal contains more distinguishable cues in the high frequency region. For this purpose, the MFCC and IMFCC features are extracted for various frequency ranges and the results of replay SSD systems are compared. This chapter also reports the best frequency range for the magnitude-based feature for replay SSD system.

5.1 High Frequency Analysis

From chapter 2, the mathematical model of the replay speech is given as:

$$r(t) = s(t) * mp(t) * pr(t) * ms(t) * pa(t), \tag{5.1}$$

where $mp(t), ms(t), pr(t)$ and $pa(t)$ represents impulse response of microphone, multimedia speaker, place of recording and place of replay attack, respectively. Let, $n(t) = mp(t) * pr(t) * ms(t) * pa(t)$, hence,

$$r(t) = s(t) * n(t), \tag{5.2}$$

where, $r(t)$, $s(t)$ and $n(t)$ represent replay speech, genuine speech and replay noise, respectively. Taking Fourier transform and its absolute value on both the sides, we get,

$$|R(\omega)| = |S(\omega)||N(\omega)|, \tag{5.3}$$

where $R(\omega), S(\omega)$ and $N(\omega)$ represents frequency-domain representation of replay speech, genuine speech and replay noise. From Eq. (5.3), we can say that, replay speech is noise affected speech over entire frequency range. The speech signal contains more lower frequencies (<6 kHz) than the higher frequencies (>6 kHz). Due to which in lower frequencies speech is dominant, while in higher frequencies noise is more dominant. The spectrograms from 6 kHz to 8 kHz frequency range are shown in Figure 5.1.

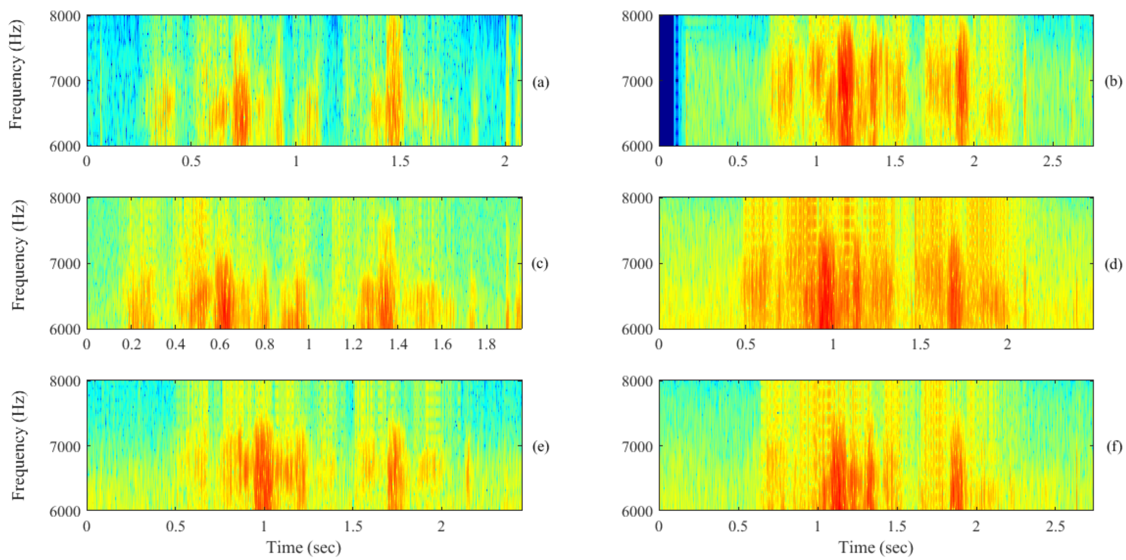


Figure 5.1: Spectrogram of (a-b) two sessions of genuine speech, (c) replay speech recorded in bedroom, (d) replay speech recorded in cafeteria, (e) replay speech recorded in office, and (f) replay speech recorded in balcony.

5.1.1 Mel Frequency Cepstral Coefficients (MFCC)

The genuine and replay speech are highly discriminative in high-frequency ranges due to dominating noise in the high-frequency range. To verify this, the MFCCs are extracted for various frequency ranges. Figure 5.2 shows the block diagram for MFCC feature extraction. The replay speech contains noise and that is the discriminative factor. Figure 5.3 shows the step-by-step output of high frequency MFCC extraction of the noisy frame for 12 coefficient using 15 filterbank.

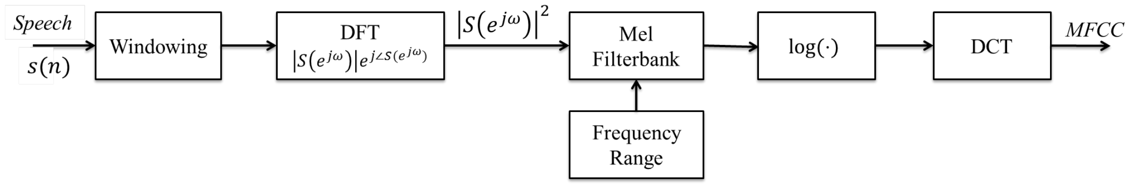


Figure 5.2: Block diagram for MFCC feature extraction.

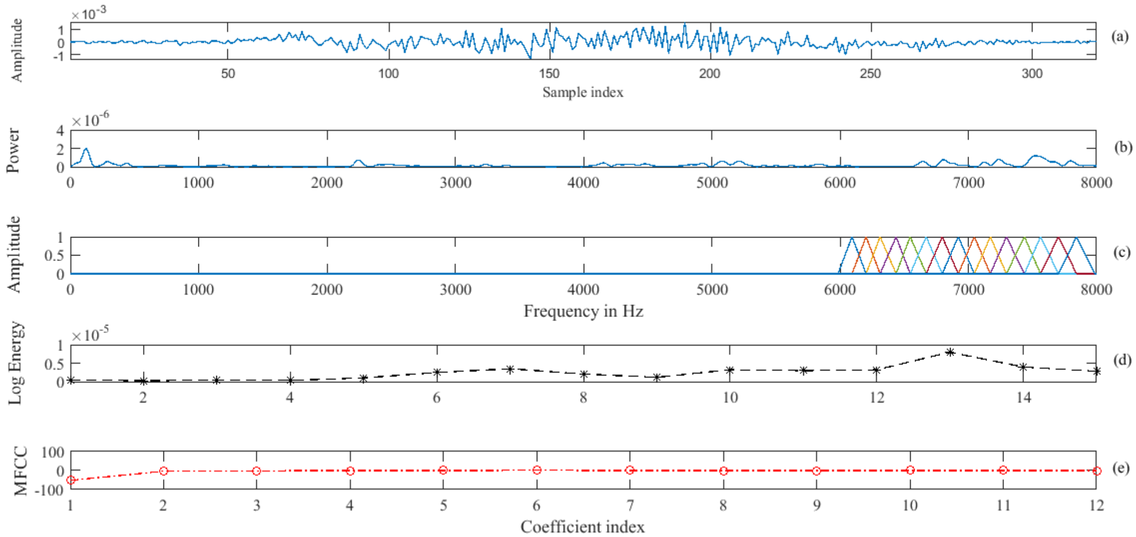


Figure 5.3: (a) Speech frame, (b) power spectrum of frame, (c) Mel filterbank for 6-8 kHz range, (d) log-energies, and (e) MFCC.

5.1.2 Inverted Mel Frequency Cepstral Coefficients (IMFCC)

The MFCC features are widely used in various speech applications. In replay spoof detection, the higher frequencies have more important than lower frequencies. Thus, it is expected that inverting Mel filterbank will give better results. Figure 5.4 shows the block diagram of IMFCC extraction.

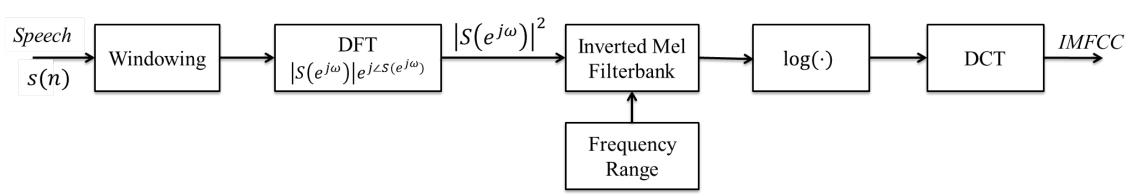


Figure 5.4: Block diagram for IMFCC extraction.

The process of IMFCC extraction is same as that of MFCC extraction except that the Mel filterbank is inverted. Figure 5.5 shows the step–by-step output of

IMFCC extraction process of the noisy frame.

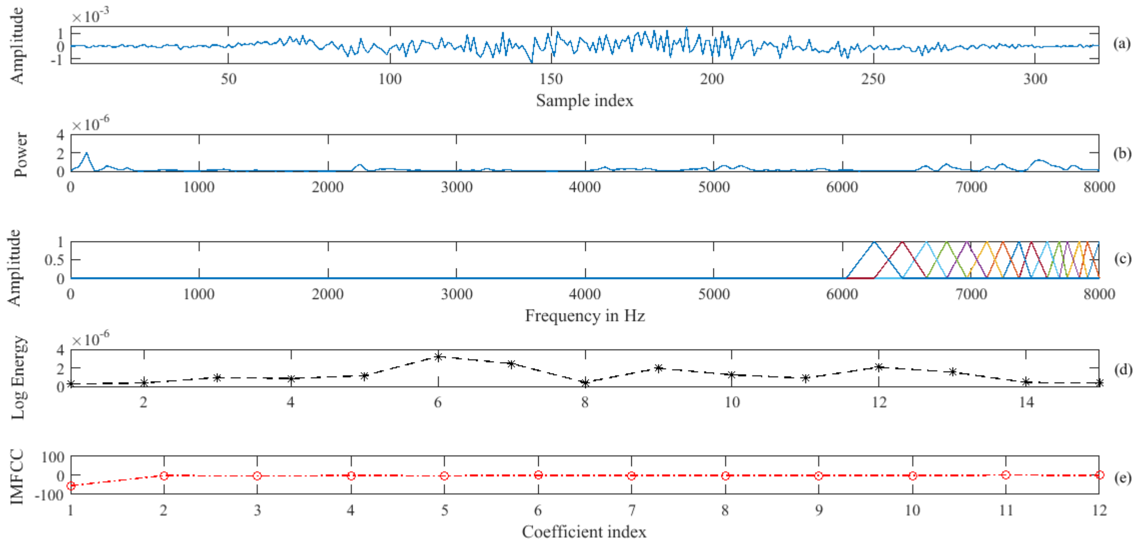


Figure 5.5: (a) Speech frame, (b) power spectrum of frame, (c) inverted Mel filter-bank for 6-8 kHz range, (d) log-energies, and (e) IMFCC.

5.2 Feature Parameterization

The 12 number of MFCC along with their delta and delta-delta coefficients are used with different frequency hop size. Total 40 Mel triangular filters (inverted Mel triangular filter for IMFCC) along with the Hamming window of 25 ms duration and 10 ms shift are used to extract MFCC predominantly from high frequency region. The MFCC and IMFCC features are extracted by using voicebox toolbox [53]. On classifier side, 512 components GMM is used.

5.3 Experiments and Results

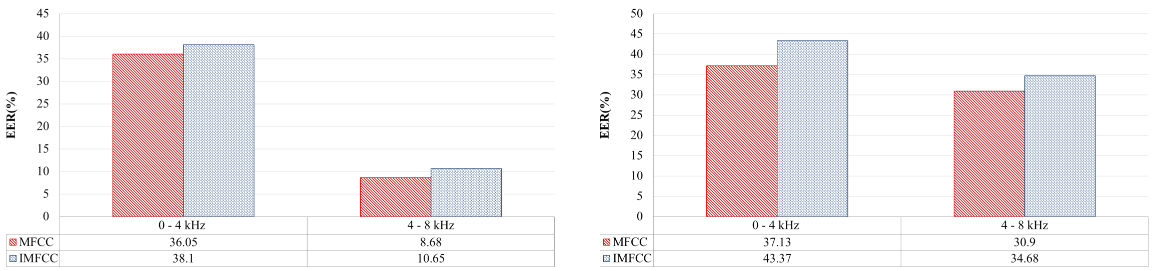
Table 5.1 shows results of replay SSD system on the development and evaluation set for entire frequency range, i.e., 0-8 kHz range.

Table 5.1: Results on development and evaluation set for entire frequency range

Feature Set	EER (%)	
	Development	**Evaluation**
MFCC	17.47	38.29
IMFCC	13.26	33.22

It is observed that the IMFCC feature set performs better than the MFCC feature set as expected in chapter discussion.

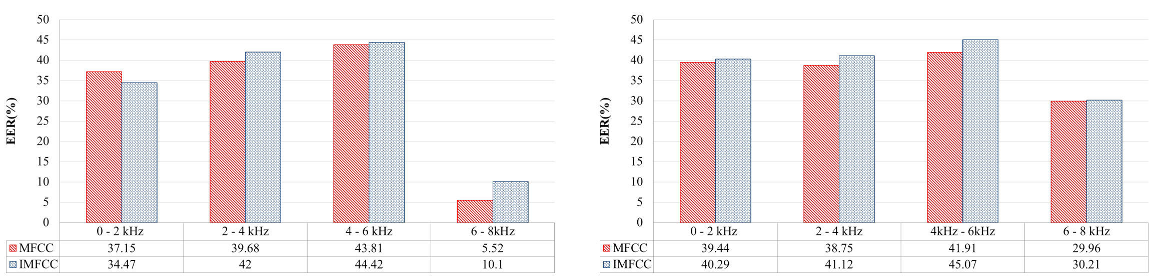
To verify the importance of the higher frequencies for replay SSD task, the entire frequency range is divided into parts with hop size of 4 kHz. Figure 5.6 shows the bar chart representation of results on development set and evaluation set for 0-4 kHz and 4-8 kHz ranges. From Figure 5.6, it is observed that the 4-8 kHz range is more distinguishable than 0-4 kHz range.



(a) (b)

Figure 5.6: Bar chart representation of results on (a) development set and (b) evaluation set with frequency hop size of 4 kHz.

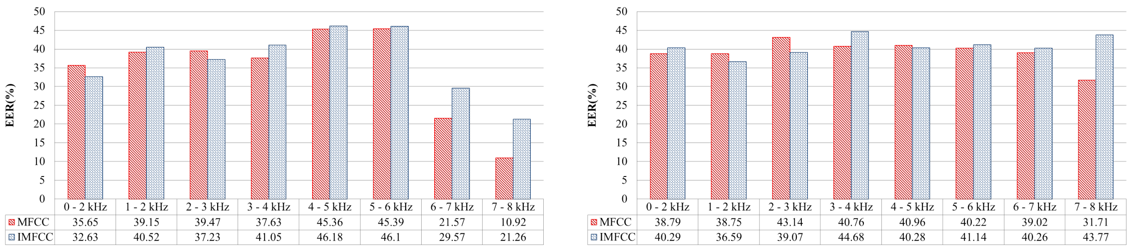
However, the speech signal is more dominant in 0-6 kHz range and hence in next set of experiments, similar kind of analysis is done for frequency hop size of 2 kHz and 1 kHz. Figure 5.7 and Figure 5.8 shows the bar chart representation of results on development set and evaluation set for hop size of 2 kHz and 1 kHz, respectively.



(a) (b)

Figure 5.7: Bar chart representation of results on (a) development set, and (b) evaluation set with frequency hop size of 2 kHz.

From Figure 5.7 and Figure 5.8 it can be observed that the 6-8 kHz range gives the better results compared to other frequency ranges, and in that also 7-8 kHz gives good results compared to 6-7 kHz. Hence, empirically it is proved that the high frequency region contain more discriminative acoustic cues than lower frequency region for the replay SSD task.



(a) (b)

Figure 5.8: Bar chart representation of results on (a) development set, and (b) evaluation set with frequency hop size of 1 kHz.

5.4 Chapter Summary

In this chapter, the frequency analysis of genuine *vs.* replay speech is carried out. It is observed that the replay speech magnitude spectrum is more distinguishable at higher frequencies (6-8 kHz) than the genuine speech magnitude spectrum. The significance of high frequency regions in replay spoof detection is empirically proved using MFCC and IMFCC feature set for different frequency ranges. In the next chapter, the novel feature set based on the spectral root homomorphic processing is discussed.

CHAPTER 6

Novel Spectral Root Cepstral Features

In this chapter, the source-filter model of the natural speech production is exploited by modeling the genuine speech signal as the convolution of excitation source (glottal airflow) and system (vocal tract) impulse response [7]. These convolutionally combined signals cannot be distinctly observed in the spectral-domain. In addition, the speech signal gets convolved with the other system responses, such as transmission channel or by the flawed recording device. In the context of replay spoof modeling, the spoofed speech signal can be modeled as a convolution of the genuine speech with the other additional elements, such as acoustic effects introduced by the recording device, recording environmental conditions, playback device and acoustics of the environment where the attack takes place. In this chapter, the newly proposed feature sets, namely, MSRCC and PSRCC are discussed for SSD system. In these feature sets, an effect of additional elements in replayed speech is over entire quefrency-domain, which is helpful in SSD task. In addition, these feature sets, contains very high complementary information as their score-level fusion gives the significant improvement in performance as compared to the standalone MSRCC or PSRCC feature sets. The proposed feature sets perform significantly better than the baseline CQCC-GMM SSD system.

6.1 Speech Signal Modeling

Let $s(n)$ be the genuine speech signal that can be modeled as a convolution of glottal airflow, $p(n)$ and vocal tract impulse response, $h(n)$ [3, 54].

$$s(n) = p(n) * h(n), \tag{6.1}$$

where $*$ denotes the convolution. Given genuine speech, $s(n)$, replay speech signal can be modeled as [3, 16]:

$$r(n) = s(n) * mp(n) * pr(n) * ms(n) * pa(n), \tag{6.2}$$

where $mp(n), ms(n), pr(n)$ and $pa(n)$ represents impulse response of microphone, multimedia speaker, place of recording and place of attack, respectively. Furthermore, Eq. (6.2) can be simplified as [3]:

$$r(n) = s(n) * N(n), \tag{6.3}$$

where $N(n) = mp(n) * pr(n) * ms(n) * pa(n)$. To detect these extra convolved elements, the homomorphic signal processing techniques can be applied. Generally, there are two homomorphic techniques, namely, the system that transfers convolutional vector space into additive vector space [55], and the system which maps convolutional vector space into another convolutionally combined vector space [56]. The purpose of both the homomorphism is to separate the convolutionally combined signals by compressing the impulse responses with respect to the impulse train of the glottal pulse.

6.2 Logarithmic Cepstrum *vs.* Spectral Root Cepstrum

To evaluate the cepstrum, the convolutionally combined signals, $s(n) = p(n) * h(n)$ maps to additively combined signals, $\hat{s}(n) = \hat{p}(n) + \hat{h}(n)$, where $(\hat{\cdot})$ indicates logarithmic cepstrum, so that the effect of impulse train $p(n)$ and impulse response of the system $h(n)$ can be observed distinctly. This transformation should take place such that $\hat{p}(n)$ remains the train of pulses with similar duration as $p(n)$, but $\hat{h}(n)$ is more time-limited than $h(n)$. The cepstrum $\hat{s}(n)$ of the signal $s(n)$ can be obtained by inverse Z transform of *log* of the Z transform of the signal. Z-transform of the signal followed by *log* ensures the transformation in additive vector space. Hence, cepstrum of the replayed speech can be expressed as [3]:

$$\hat{r}(n) = \hat{s}(n) + \hat{N}(n). \tag{6.4}$$

To obtain spectral root cepstrum, convolutionally combined signal, $s(n) = p(n) * h(n)$ is mapped to another convolutionally combined signal, $\check{s}(n) = \check{p}(n) * \check{h}(n)$, where $(\check{\cdot})$ indicates spectral root cepstrum, such that the new convolutionally combined vectors are more easily separable. In spectral root cepstrum, logarithmic operator during cepstrum computation is replaced by exponent γ. Z transform maps the convolutional vector space into multiplicative vector space to give $S(z) = P(z) \cdot H(z)$. Then,

$$\begin{aligned} \check{S}(z) &= [P(z) \cdot H(z)]^{\gamma}, \\ \check{S}(z) &= P^{\gamma}(z) \cdot H^{\gamma}(z), \\ \therefore \check{S}(z) &= \check{P}(z) \cdot \check{H}(z). \end{aligned} \tag{6.5}$$

For above set of equations [3], there is a one-to-one mapping between the time-domain vectors and the corresponding Z-domain vectors. In addition, there is an implicit assumption that $s(n)$ is a real and stable sequence. Taking inverse Z transform of Eq. (6.5), we get [3],

$$\check{s}(n) = \check{p}(n) * \check{h}(n), \tag{6.6}$$

where $\check{s}(n)$ is known to be spectral root cepstrum.

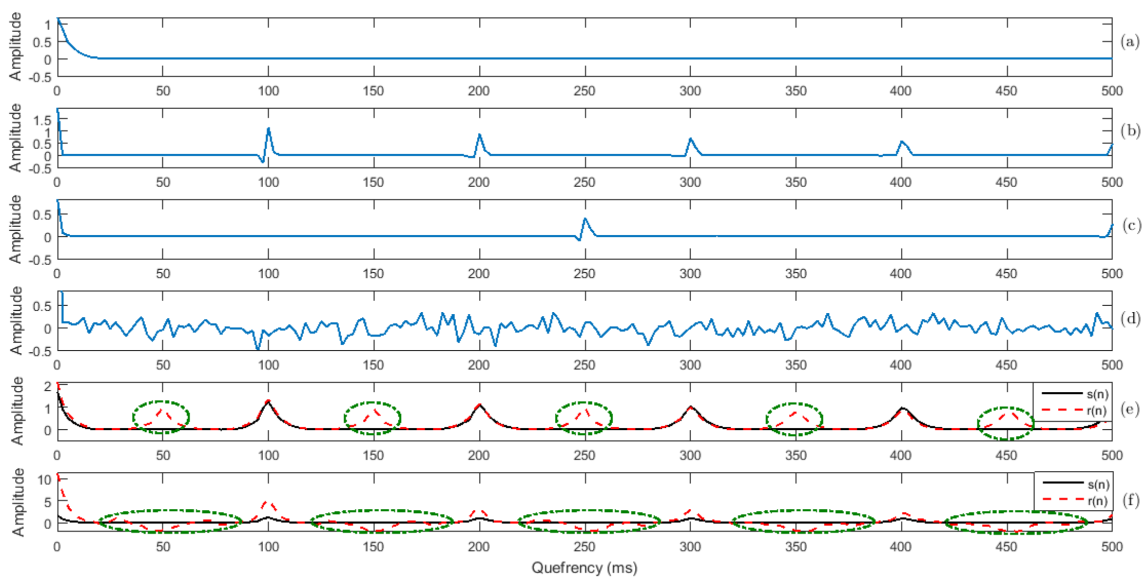


Figure 6.1: Spectral root cepstrum of synthetic (a) $h(n)$, (b) $p(n)$, (c) $N(n)$ (impulsive), (c) $N(n)$ (white Gaussian noise), (e) $s(n)$ and $r(n)$ for impulsive noise, and (f) $s(n)$ and $r(n)$ for white Gaussian noise. After [3] (*highlighted portion indicates effect of* $N(n)$ *component*).

Since $\hat{S}(z) = \log(S(z))$ and $\check{S}(z) = (S(z))^{\gamma}$, $\check{S}(z)$ is related to $\hat{S}(z)$ by [3]:

$$\check{S}(z) = e^{\gamma \cdot \hat{S}(z)} = \sum_{n=0}^{-\infty} \frac{(\gamma \hat{S}(z))^n}{n!}, \tag{6.7}$$

$$\therefore \check{s}(n) = \delta(n) + \gamma\hat{s}(n) + \frac{\gamma^2}{2}\hat{s}(n) * \hat{s}(n) + \frac{\gamma^3}{3}\hat{s}(n) * \hat{s}(n) * \hat{s}(n) + \cdots . \quad (6.8)$$

From Eq. (6.8), it can be observed that even though $s(n)$ is of limited duration, $\check{s}(n)$ will be having infinite duration. For genuine speech signal $s(n)$ and spoofed speech signal $r(n)$, spectral root cepstrum can be given as [3]:

$$\check{s}(n) = \check{p}(n) * \check{h}(n), \quad (6.9)$$

$$\check{r}(n) = \check{s}(n) * \check{N}(n) = \check{p}(n) * \check{h}(n) * \check{N}(n). \quad (6.10)$$

From above set of equations, the effect of $N(n)$ in logarithmic *vs.* spectral root cepstrum can be analyze. As shown in Eq. (6.4), $\hat{N}(n)$ is additive to $\hat{s}(n)$, its effect in $\hat{r}(n)$ will have limited support. While in Eq. (6.10), $\check{N}(n)$ is convolved with $\check{s}(n)$. Hence, effect of $\check{N}(n)$ spreads across entire quefrency-domain. This property can be validated from Eq. (6.8), as it shows that spectral root cepstrum is the linear combination of convolution of logarithmic cepstrum. Hence, additively combined components of logarithmic cepstrum are squeezed in spectral root cepstrum due to convolution. The same analysis is demonstrated in Figure 6.1. For verification, four synthetic signals are chosen, $h(n)$ (Figure 6.1a), $p(n)$ (Figure 6.1b), $N(n)$ impulsive (Figure 6.1c), and $N(n)$ white Gaussian noise (Figure 6.1d). The two signals, i.e., $s(n) = h(n) * p(n)$ and $r(n) = h(n) * p(n) * N(n)$ are depicted in each Figure 6.1e *($N(n)$ is impulsive)* and Figure 6.1f *($N(n)$ is white Gaussian noise)*. The effect of $N(n)$ is present across the quefrency-domain because of its convolution with $s(n)$. However, in logarithmic cepstrum, the effect of $N(n)$ is concentrated in particular region because of addition of various cepstra (Eq. (6.4)). Hence, spoofed speech can be discriminated well in spectral root cepstrum than the logarithmic cepstrum.

6.3 Proposed Feature Extraction

It has been observed that the average auditory nerve firing rate shows an overshoot at the onset of an input signal. In addition, studies shows that the human auditory system appears to focus on the onset of incoming power envelope rather than the falling edge of the same power envelope [57]. Thus, the human auditory system can be modeled by the functional relationship between the onset firing rate of auditory neurons and Sound Pressure Level (SPL). Studies have shown that the given relationship can be approximated by power law nonlinearity [58, 59]. Another advantage of this nonlinearity is that its asymptotic response to the lower amplitude signals approaches to zero rather than negative infinity unlike in MFCC [60]. In addition, it approximates the "psychophysical transfer function" which relates the physical intensity of sensation to perceived intensity using direct magnitude estimation procedures [61]. Empirically, it is found that various values of gamma gives better speech recognition accuracy for different noise models [62].

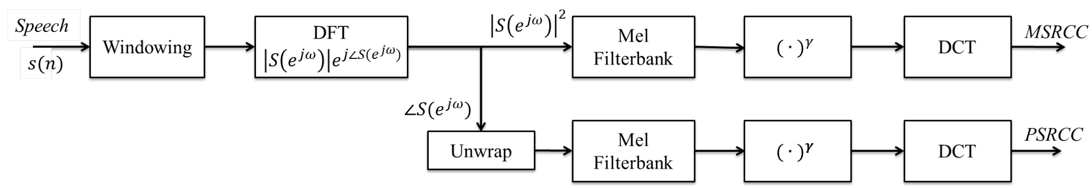


Figure 6.2: Functional block diagram of proposed MSRCC and PSRCC feature extraction methodology. After [3].

The proposed feature sets uses power law nonlinearity for the replay SSD task. Empirically, it has been observed that it gives better classification accuracy for γ=$-\frac{1}{7}$, as it may detect relevant variability due to $N(n)$ in a spoofed speech.

6.3.1 Magnitude-based Spectral Root Cepstral Coefficients (MSRCC)

The MSRCCs of a time-domain signal $s(n)$ can be obtained by the inverse transformation of spectral energy coefficients raised to a certain exponent γ. DCT [52] is used to take inverse transformation as it transforms the N real coefficients onto q real independent cepstral coefficients such that $q \ll N$, which extracts the significant information. Mathematical expression of the proposed feature set is given as [3]:

$$MSRCC(q) = \sum_{m=1}^{M} (MFM(m))^{\gamma} \, cos \left[\frac{q(m - \frac{1}{2})\pi}{M} \right], \tag{6.11}$$

where the Mel Frequency Magnitude (MFM) spectrum is defined as [3]:

$$MFM(m) = \sum_{k=1}^{K} |S(k)|^2 H_m(k), \tag{6.12}$$

where $S(k)$ is the k-point DFT of signal $s(n)$, $H_m(k)$ is the triangular weighting-shaped function for the m^{th} Mel scaled bandpass filter.

6.3.2 Phase-based Spectral Root Cepstral Coefficients (PSRCC)

Information contained in the phase part of STFT is taken into account by the PSRCC. As shown in Figure 6.2, in the development of PSRCC, spectral energy coefficients are replaced by the unwrapped phase. Mathematically [3],

$$PSRCC(q) = \sum_{m=1}^{M} (MFP(m))^{\gamma} cos \left[\frac{q(m-\frac{1}{2})\pi}{M} \right], \tag{6.13}$$

where the Mel Frequency Phase (MFP) spectrum is defined as [3]:

$$MFP(m) = \sum_{k=1}^{K} \measuredangle S(k) H_m(k), \tag{6.14}$$

where k is the DFT index.

6.4 Feature Parameterization and Results

6.4.1 Effect of Gamma Values

In the last chapter, empirically it is proved that for magnitude-based features the higher frequencies contains more cues for replay detection *(since in this range noise magnitude spectrum is more dominant than the speech signal)*. The MSRCC features are extracted for the frequency range of 6-8 kHz and PSRCC features for the entire auditory frequency range. The *13*-dimensional (D) static MSRCC and PSRCC features along with Δ and $\Delta\Delta$ coefficients are used to get *39*-D feature vector. Total 40 triangular filters along with the Hamming window of 20 *ms* duration and 50 % overlap are used in both the feature extraction process. The experiment is performed for various gamma values. Figure 6.3 shows the effect of various gamma values on the SSD system. Empirically, it is found that $\gamma=-\frac{1}{7}$ is the best choice for both the feature sets.

6.4.2 Effect of Dimension of Feature Vector

In this experiment, the effect of dimension of MSRCC and PSRCC feature vectors on SSD is analyzed. In particular, the SSD system with feature dimensions ranging from 21 to 57 *(static*+Δ+$\Delta\Delta$*)* and 512 Gaussian mixture components is studied using the best choice of gamma from the last experiment ($\gamma=-\frac{1}{7}$). Figure 6.4 shows the effect of feature dimension on SSD system. We observe that *39*-D (13-static+13-Δ+13-$\Delta\Delta$) features are sufficient and give relatively best results.

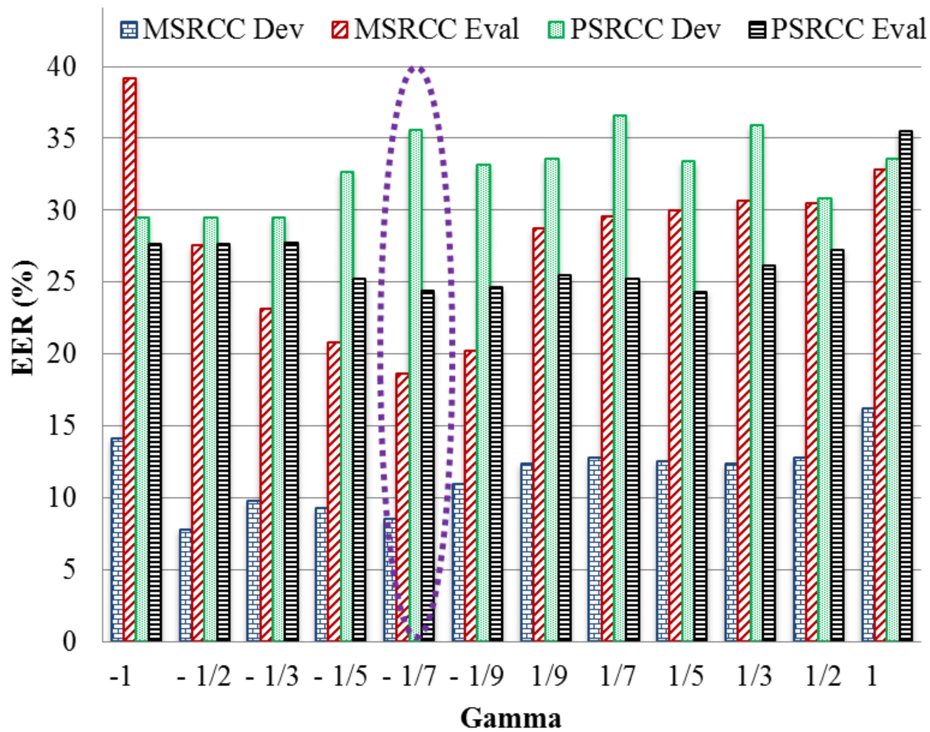


Figure 6.3: Bar chart representation depicting the effectiveness of γ values. After [3] *(Dotted ellipse above indicate relatively better result for $\gamma = -1/7$).*

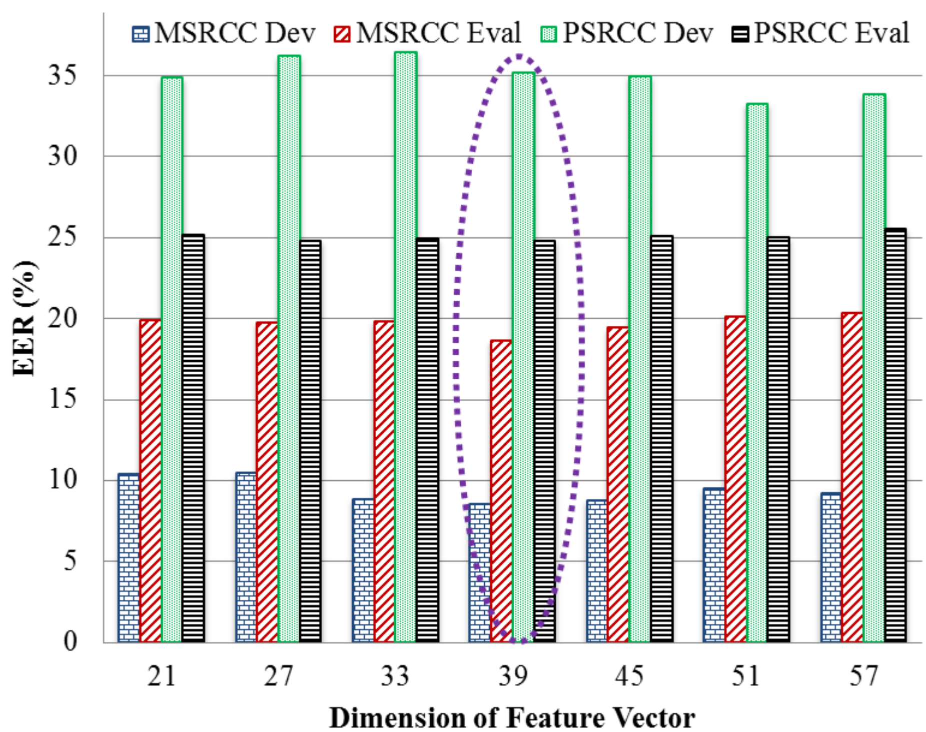


Figure 6.4: Bar chart representation depicting the effectiveness of feature dimension. After [3]. *(Dotted ellipse above indicate relatively better results for 39-D feature vector).*

6.4.3 Effect of Number of Gaussian Components

Furthermore, the effect of the number of Gaussian mixture components is examined. GMM is trained using the training set only. It is found that the 256 and 512 Gaussian components gave the best results for *39*-D PSRCC and MSRCC feature sets, respectively. The Figure 6.5 shows the effectiveness of the number of Gaussian components on the SSD system.

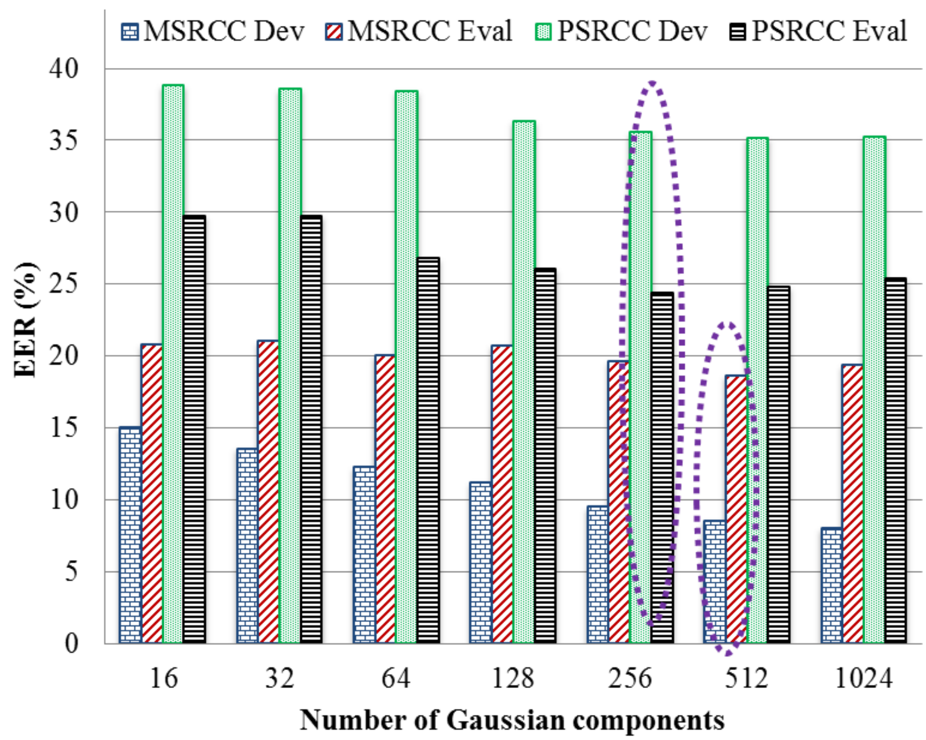


Figure 6.5: Bar chart representation depicting the effectiveness of number of Gaussian components in GMM. After [3]. *(Dotted ellipse above indicate relatively better results for 256 and 512 components for PSRCC and MSRCC, respectively).*

6.4.4 Final Replay Spoof Detection System

In GMM classification, two models are used to represent natural and spoofed speech classes. GMM for each class has 512 and 256 components for the MSRCC and PSRCC, respectively. GMM is trained using 30 iterations of Expectation Maximization (EM) algorithm. To explore possible complementary information to that of magnitude part alone, the score-level fusion of MSRCC and PSRCC is investigated. The results of the SSD on development and evaluation sets are presented in Table 6.1, where '+' indicates the score-level fusion. Figure 6.6 shows all the DET curves for development and evaluation sets of the ASV spoof 2017 database.

Table 6.1: Results for development and evaluation set. After [3]

System	EER (%)	
	Development	Evaluation
CQCC	12.11	29.18
MFCC	11.21	31.30
MSRCC	8.53	18.61
PSRCC	35.53	24.35
MSRCC + PSRCC	6.58	**10.65**

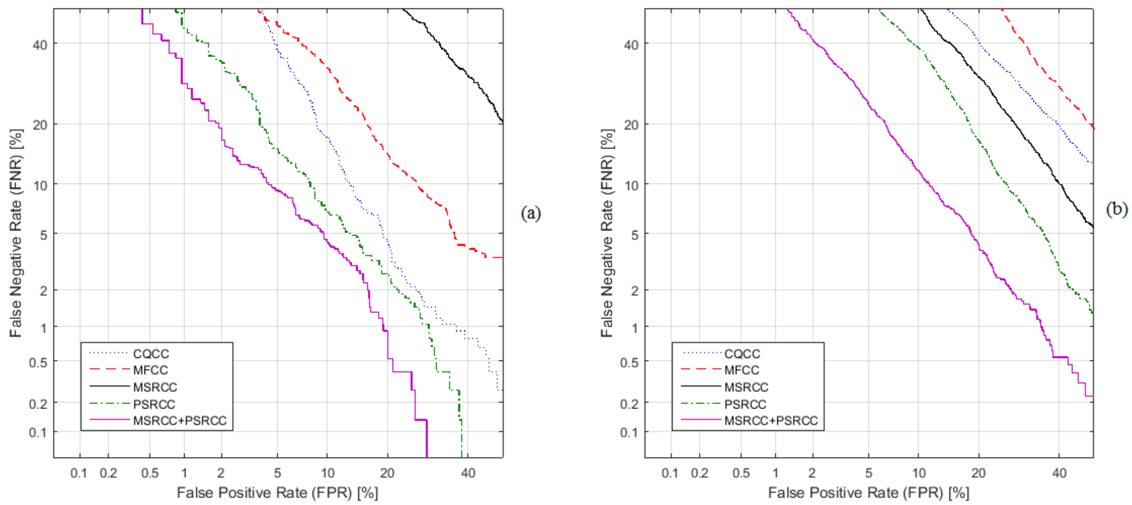


Figure 6.6: DET curves for (a) development set, and (b) evaluation set of ASV spoof 2017 challange database. After [3].

6.5 Chapter Summary

In this chapter, newly proposed MSRCC and PSRCC feature sets to detect replay spoofed speech are discussed. The proposed feature sets uses the spectral root cepstrum to characterize the natural and replayed speech. The individual system is developed and then fused at the score-level using the GMM classifier. These feature sets contains significant complementary information resulting in improved system performance. The proposed feature set performs better than the ASV spoof 2017 challenge baseline CQCC system. However, for the practical applications, higher security is more important than the higher user convenience. In the next chapter, the significance of TEO phase for replay SSD task is explored which provides higher security information.

CHAPTER 7

Teager Energy Operator (TEO) Phase Features

Several past studies have ignored phase information for various speech processing applications. In this chapter, the excitation source-like feature set, namely, TEO phase is explored and its significance in the replay spoof detection task. The TEO phase feature set was originally proposed for speaker recognition task [63]. The TEO phase captures the excitation source-related information, which is complementary to speaker-specific information obtained through spectral features, such as CQCC, MFCC, etc. [63]. The advantage of using TEO phase is that it does not require pre-processing operations, such as framing, windowing, pre-emphasis, etc. and hence, reducing computational complexity. In addition, few experiments are carried out to capture possible complementary information present in magnitude-based features, namely, CQCC [64], MFCC [42], and LFCC [65].

7.1 Teager Energy Operator (TEO) Phase

Various conventional features, such as MFCC, LPCC assumes that the speech production mechanism is linear in which the airflow propagation through vocal tract is linear planar wave. However, the concomitant vortices are dispersed over entire vocal tract area and the airflow is separated and hence, the assumption of linearity may fail [66, 67]. The actual source of speech production is vortex-flow interactions, these vortex-flow interactions are nonlinear in nature. The TEO is a nonlinear energy tracking operator for signal analysis and to characterize the airflow properties in vocal tract [66]. Considering the fact that energy required for producing an acoustical signal (such as speech) is dependent on its frequency as well as amplitude, Kaiser developed a TEO operator, $\psi(n)$, for discrete-time signal $s(n)$ as [4, 68]:

$$\psi(n) = \psi\{s(n)\} = s^2(n) - s(n+1)s(n-1). \tag{7.1}$$

Around GCIs, the TEO profile gives higher energy value. Motivated by a study reported in [69], the authors in [63] used phase of an analytic signal obtained from TEO profile of speech frame. The analytic signal $\psi_a(n)$ for TEO profile is given by [4]:

$$\psi_a(n) = \psi(n) + j\hat{\psi}(n), \tag{7.2}$$

where $\hat{\psi}(n)$ is a Hilbert transform of $\psi(n)$. The Hilbert transform produce the phase shift of 90^o for every frequency component and can be computed as follows [4]:

$$\hat{\psi}(n) = \mathcal{F}^{-1}\left(\hat{\Psi}(\omega)\right), \tag{7.3}$$

where $\mathcal{F}^{-1}$ is inverse Fourier transform and $\hat{\Psi}(\omega)$ is Fourier transform of $\hat{\psi}(n)$ given as [4]:

$$\hat{\Psi}(\omega) = \begin{cases} -j\Psi(\omega), & if \quad 0 \leq \omega < \pi, \\ j\Psi(\omega), & if \quad -\pi \leq \omega < 0, \end{cases} \tag{7.4}$$

where $\Psi(\omega)$ denotes Fourier transform of the TEO profile $\psi(n)$. The amplitude envelope of analytic signal, also known as Hilbert Envelope (HE) is given by [4]:

$$a_e(n) = \sqrt{\psi^2(n) + \hat{\psi}^2(n)}. \tag{7.5}$$

The TEO phase is cosine of the phase of analytical signal $\psi_a(n)$ and computed as [4]:

$$\phi_\psi(n) = \cos\left(\measuredangle\psi_a(n)\right) = \frac{\psi(n)}{a_e}. \tag{7.6}$$

where $\phi_\psi(n)$ denotes the TEO phase. The Figure 7.1 shows the voiced segment of speech signal, its TEO profile, the Hilbert transform of TEO profile, Hilbert envelope and TEO phase for genuine (Panel I) and similar analysis for corresponding replay speech (Panel II). The Figure 7.2 shows the similar analysis for speech segment containing silence region followed by voiced region for genuine (Panel I) and replay speech (Panel II). From Figure 7.1, it can be observed that the TEO phase plot of the replay speech (panel II) is more fluctuating compared to the genuine speech (panel I) signal in case of voice speech segment. From Figure 7.2, it can be noticed that the genuine speech (panel I) signal containing silence region gives almost zero TEO phase values for silence region, unlike replay speech (panel II) signal which gives significant TEO phase values in silence region (because small bumps present in Hilbert envelope of silence region).

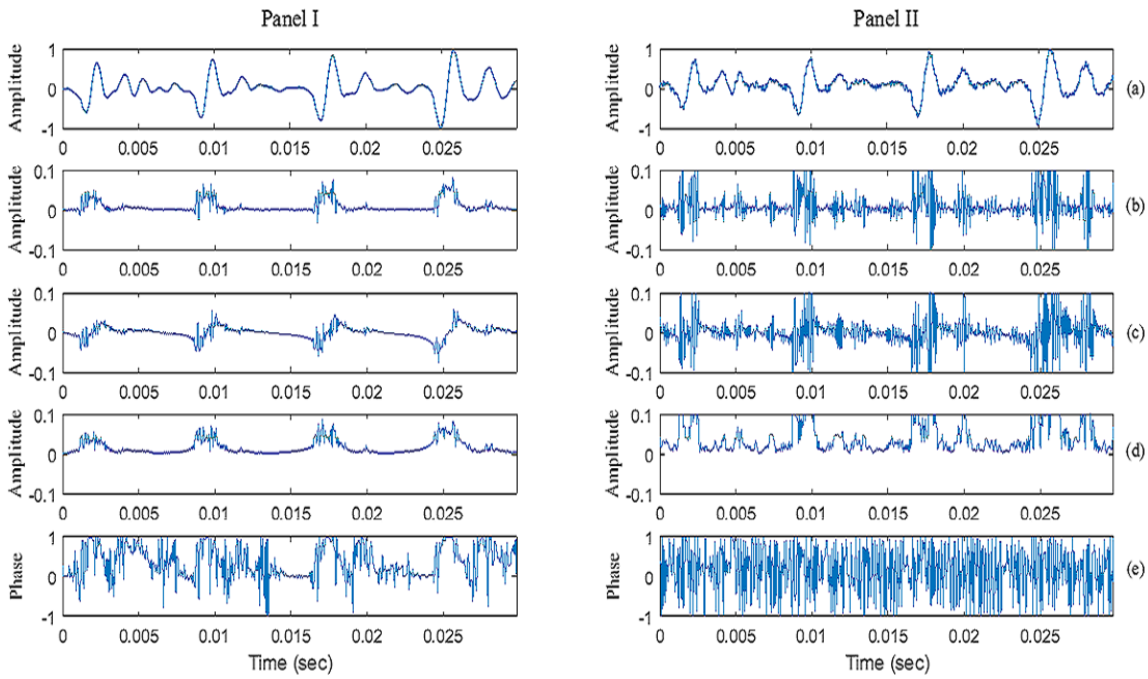


Figure 7.1: (a) Voiced speech segment, (b) TEO profile, (c) Hilbert transform, (d) Hilbert envelope, and (e) TEO phase (Panel I: Genuine speech segment, Panel II: Corresponding replay speech segment). After [4].

The another observation is that although TEO profile indicates energy, it can have negative values (as can be observed from Eq. (7.1)) and have higher energy values when vocal tract gets sudden impulse-like excitation. From Figure 7.1 and Figure 7.2, it can be observed that for genuine speech, TEO profile gives higher values near GCIs, however, for replay speech, TEO profile gives higher values around GCIs as well as other locations. This may be due to the noise present in replay speech signal which contribute to running estimate of energy. It is also observed that the TEO phase has better correlation with input speech signal.

From Figure 7.2, it is clear that for silence region of genuine speech TEO profile has approximately zero energy and hence, HE and TEO phase also have zero energies. However, in the replay speech, presence of some noisy samples results in spurious TEO values and hence, HE and TEO phase have non-zero energies. It is also observed that the energy values at GCIs for replay speech gets magnified compared to the genuine speech signal, this may be due to fact that replay speech signal is noisy genuine speech signal (replay can be modeled as convolution of genuine speech signal with impulse response of intermediate devices, impulse response of recording and playback environment).

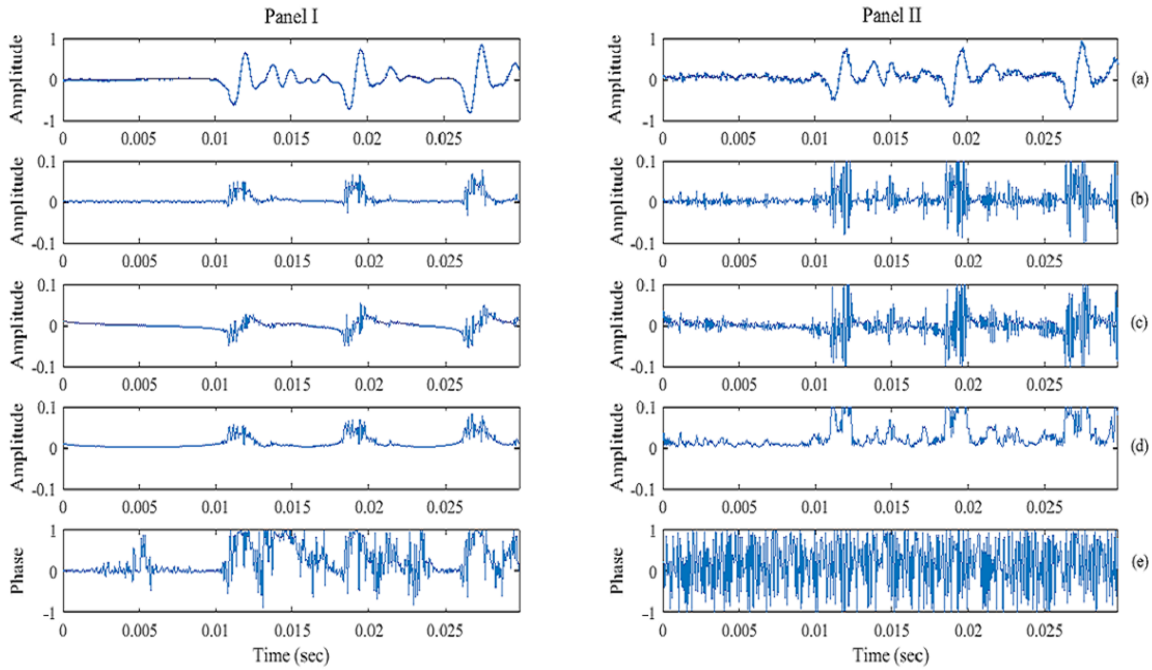


Figure 7.2: (a) Speech signal having silence region followed by voiced segment, (b) TEO profile, (c) Hilbert transform, (d) Hilbert envelope, and (e) TEO phase. (Panel I: Genuine speech segment, Panel II: Corresponding replay speech segment). After [4].

From these observations, it can be hypothesized that the TEO phase information has potential for replay spoof detection task. This is explored in the next section.

7.2 Feature Extraction

Figure 7.3 shows the schematic diagram to estimate the TEO phase feature set. Here, first the TEO profile of the input speech signal is computed using Eq. (7.1). The Hilbert envelope of the TEO profile is computed from analytic signal of TEO profile using Eq. (7.5). The feature vector is formed by taking B blocks (each of N_d samples) of TEO Phase with some shift at the GCI. However, this requires estimation of location of GCIs. Figure 7.1 and Figure 7.2 shows that the TEO profile is blunted and hence, the better singularity detection algorithm (for GCIs estimation) is required. The multiscale edge detection can be done using Canny edge detector which is equivalent to wavelet modulus maxima using Gaussian kernel.

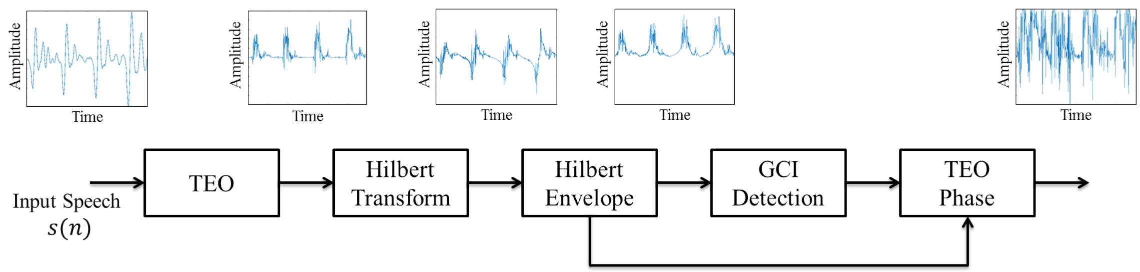


Figure 7.3: Functional block diagram to extract TEO phase feature set. After [4].

For singularity detection, wavelet analysis is used. To do this, first local fluctuations in Hilbert envelope needs to removed. To get rid of these fluctuations, local mean smoothing followed by its wavelet transform of Hilbert envelope is done. The wavelet transform of a signal can be expressed as multiscale differential operator [70]. In [71], it is reported that all the singularities present in signal can be detected using wavelet transform modulus maxima at finer scales. This property of signal is used for GCI detection in TEO phase feature extraction. The derivatives of the Gaussians are widely used in numerical computations to make sure all the maxima line propagate up to the fine scales (pp.177-178, [70]). As TEO profile is calculated for entire input speech signal, it avoids voiced/unvoiced detection, pre-emphasis, framing and have less computation cost.

7.3 Feature Parameterization

System S_1 built with TEO phase feature set. The 6 blocks each of 40 samples with one sample shift of TEO Phase at the GCI is taken to form *40*-dimensional (D) feature vector. The GCIs are estimated using Hilbert envelope and *1*-D Canny operator. The system S_2 is built with *90*-D CQCC feature set that comprise of the zeroth coefficient, 29-static, 30-Δ, and 30-$\Delta\Delta$ coefficients. The minimum frequency set to 15 Hz and maximum frequency to 8 kHz, the number of bins per octave set to 96. System S_3 developed with *39*-D (13-static + 13-Δ + 13-$\Delta\Delta$) MFCC feature set. Total 40 triangular filters along with the Hamming window of 20 ms duration and 10 ms shift are used for the feature extraction process. System S_4 is based on LFCC. The LFCC feature set is extracted with 60 triangular filters and using frame length of 20 ms with 50 % overlap. Extracted features are appended with 60-Δ and 60-$\Delta\Delta$ coefficients resulting into *180*-D feature vector. Table 7.1 summarizes the experimental setup used for development of spoof detection system.

Table 7.1: Summary of the experimental setup of the systems *(FD: Feature Dimension)*. After [4]

System	Feature Set	FD	No. of Gaussians
S_1	TEO phase	40	256
S_2	CQCC *(Baseline)*	90	512
S_3	MFCC	39	512
S_4	LFCC	180	512

7.4 Experimental Results

The experimental results of the replay spoof detection on development and evaluation sets are given in Table 7.2, where '+' indicates the score-level fusion. From results, it can be observed that the individual TEO phase feature do not perform well. However, when they are fused with magnitude spectrum-based feature sets, the system performance improves substantially. This indicates that the TEO phase feature contains complementary information to the magnitude-based features.

Table 7.2: Results for development and evaluation sets. After [4]

System	EER (%)	
	Development	Evaluation
TEO phase	23.91	31.34
CQCC	11.89	28.92
MFCC	17.27	34.02
LFCC	10.28	16.80
TEO phase+CQCC	10.62	28.74
TEO phase+MFCC	14.56	31.28
TEO phase+LFCC	**6.57**	**15.39**

The organizers of the ASV spoof challenge provided CQCC-GMM as baseline system with an EER of 28.92 % on evaluation set of database. The standalone spoof detection system built with TEO phase, MFCC and LFCC gives a result of 31.34 %, 34.02 %, and 16.80 %, respectively, on evaluation set. When TEO phase feature set is fused with CQCC, MFCC, and LFCC, EER gets reduced by 0.18 %, 2.74 %, and 1.41 %, respectively, compared to the corresponding magnitude spectrum-based feature sets. This improvement in system performance points out that the presence of TEO phase along with magnitude information strengthens the spoof detection system. Figure 7.4 shows the DET curves for development and evaluation set of ASV Spoof 2017 Challenge database.

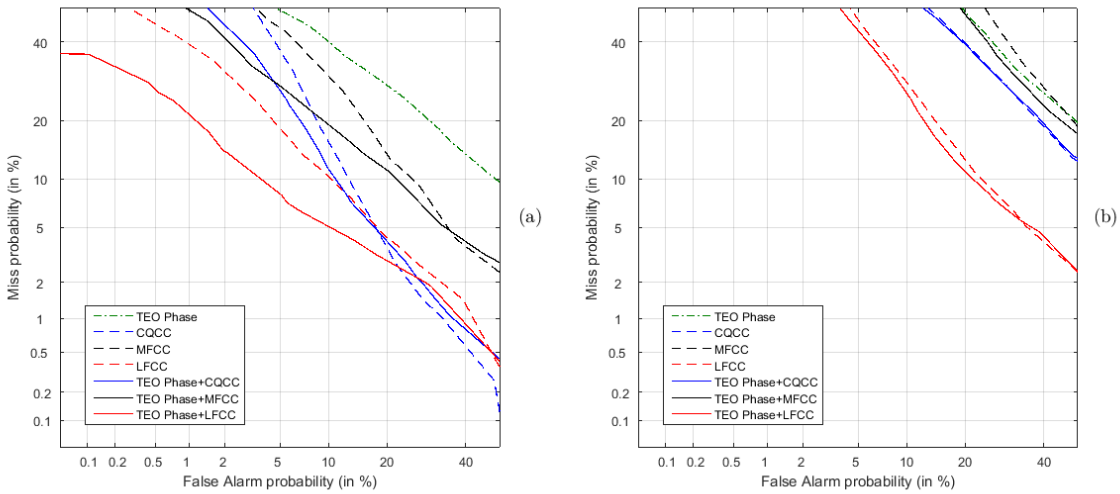


Figure 7.4: DET curves for (a) development set, and (b) evaluation set. After [4].

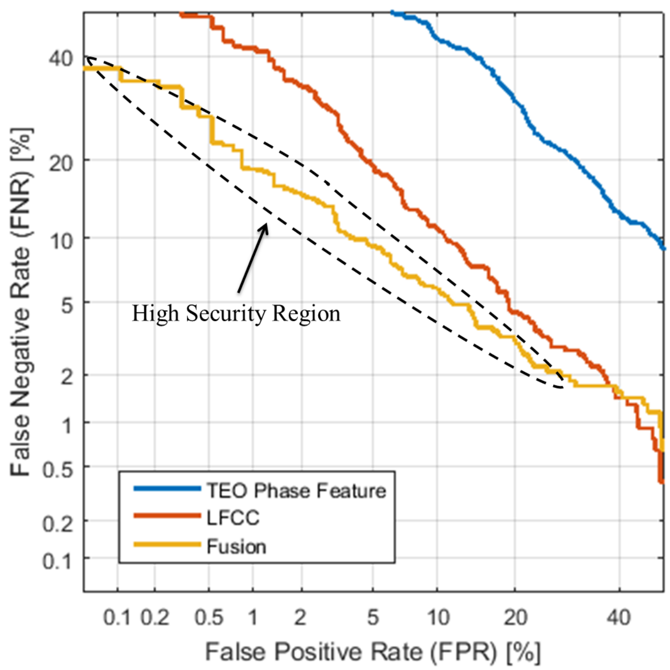


Figure 7.5: DET curves for TEO phase, LFCC and their score-level fusion for development set *(highlighted portion indicated deviation towards high security region)*. After [4].

Figure 7.5 shows the DET curve for TEO phase, LFCC and their score-level fusion for development set, similar curves were observed for CQCC, MFCC and IM-

FCC. From the DET curves, it is observed that when magnitude spectrum-based feature sets are fused with TEO phase feature set the DET curve deviates towards vertical-axis more compared to the horizontal-axis, i.e., the probability of false acceptance is less. However probability of false rejection is comparatively high. This indicates that the TEO phase feature set capture the information required for designing high security replay spoof detection system for ASV. TEO phase feature set detects the spoofed speech very efficiently and does not allow the attacker to access the ASV system easily, which is very important in practical applications.

7.5 Chapter Summary

In this chapter, the significance of TEO phase feature set is explored for replay spoof detection task. It is observed that the TEO phase plots seem to be very noisy for replay speech compared to natural speech. In this chapter, the TEO phase feature set performance is investigated with CQCC, MFCC and LFCC feature sets. It is observed that the TEO phase feature gives the complementary information to the speaker-specific information provided by CQCC, MFCC and LFCC feature sets. It is also observed that the TEO phase feature set provide an information which shift DET curve towards high-security reason than the high user convenience region, indicating that TEO phase efficiently detects replayed speech (i.e., better performance w.r.t. anti-spoofing). The next chapter summarizes entire work along with some future research directions of the work.

CHAPTER 8

Summary and Conclusions

8.1 Conclusions from the Thesis Work

The conventional feature sets, such as MFCC, LPCC are designed for speaker recognition task. The thesis aims to develop various features for the replay spoof detection task. In the thesis, three novel system-based cepstral feature sets, namely, EMDCC, MSRCC and PSRCC are proposed. Furthermore, a source-like feature set, namely, TEO phase is also explored for replay spoof detection problem. Some of the observations and conclusions from the thesis work are as follows:

- The EMD decomposes the input speech signal into IMFs assuming that the signal is nonlinear and nonstationary and act as a dyadic filterbank. The zero crossing rate of IMFs are approximated as a center frequencies of band-pass filters of filterbank.
- In the EMDCC feature set, the power law nonlinearity performs better than the logarithmic nonlinearity, however, it is observed that logarithmic nonlinearity captures better complementary information compared to the power law nonlinearity.
- It is observed that the replay speech magnitude spectrum is more distinguishable at higher frequencies (6-8 kHz) than the genuine speech magnitude spectrum.
- The score-level fusion of the MSRCC and PSRCC feature sets improves the system performance significantly, indicating that these feature sets contains significant complementary information.
- The PSRCC feature set is giving better results on evaluation set compared to the development set, indicating that the PSRCC feature set perform well in unseen conditions.

- It is observed that the TEO phase plots are very noisy for replay speech compared to the natural speech. For the silence region of the genuine speech signal, TEO profile has approximately zero energy and hence, zero TEO phase values. However, for replay speech signal, TEO phase has nonzero values for the silence regions.

- When magnitude spectrum-based features are fused with TEO phase feature set the DET curve shift towards vertical-axis more compared to the horizontal-axis indicating that the TEO phase feature capture the information required for designing high-security replay spoof detection system for ASV (which is all the more desirable for goal of anti-spoofing). Hence, TEO phase feature set detects the spoofed speech very efficiently and does not allow the attacker to access the ASV system easily, which is very important for practical applications.

8.2 Limitations of the Thesis Work

The limitations of the thesis work are as follows:

- In the present work, only GMM classifier is used to develop the replay SSD systems. The limitation of GMM is that the number of Gaussian mixture components is unknown.

- The feature discrimination measures of the proposed feature sets are not evaluated.

- The number of subbands obtained through the filterbank structure is same for the all utterances. However, in EMDCC feature set, the number of IMFs extracted from all the utterances may not be the same.

- In the spectral root cepstral features, the value of gamma needs to find out manually and it depends on the environmental conditions.

8.3 Future Research Directions

The possible future research directions for the work presented in the thesis are as following:

- In the present work, the GMM classifier is used to developed replay SSD system. However, the neural network-based classifier, such as CNN, DNN, BLSTM gives the better discrimination and performs better.

- Further work can be extended to evaluate the feature discrimination of each proposed set using performance metrics, such as F-ratio, J-measure, etc.
- The replay speech can be modeled as the convolution of genuine speech as channel noise. The EMD can be used to estimate the channel noise and this estimated noise can be used to develop replay SSD system.
- The spectral root cepstral features can be explored for different frequency scales, such as linear, ERB, Bark, log10, etc. and various filterbanks, such as Butterworth, rectangular, Gabor, etc.
- In future, Variable length TEO (VTEO) phase can be explored with magnitude spectrum information for better system performance.

CHAPTER A
EM Algorithm for GMM Parameter Estimation

The likelihood function for D-dimensional feature vector x and K component GMM is defined as,

$$P(x) = \sum_{k=1}^{K} w_k P_k \left(x|\mu_k, \Sigma_k\right). \tag{A.1}$$

The density is a weighted linear combination of K unimodal Gaussian densities, $P_k \left(x|\mu_k, \Sigma_k\right)$ each parameterized by a mean $D \times 1$ vector μ_k, and a $D \times D$ covariance matrix, Σ_k:

$$P(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma|^k} e^{-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)}. \tag{A.2}$$

The EM algorithm can be summarized as [72]:

Algorithm 1 EM algorithm for parameter estimation of GMM

1: Initialize w_k, μ_k, and Σ_k for $k = 1, 2, ..., K$.

2: E-Step (Finding Latent Variable)

$$z_k^i = \frac{w_k P_k(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} w_k P_k(x_i|\mu_k, \Sigma_k)}.$$

3: M-Step (Updating parameters by maximizing likelihood)

$$\begin{aligned} w_k &= \frac{\sum_{i=1}^{N} z_k^i}{N}, \\ \mu_k &= \frac{1}{\sum_{i=1}^{N} z_k^i} \sum_{i=1}^{N} z_k^i x_i, \\ \Sigma_k &= \frac{1}{\sum_{i=1}^{N} z_k^i} \sum_{i=1}^{N} z_k^i (x_i - \mu_k)(x_i - \mu_k)^t. \end{aligned}$$

4: Repeat step 2 and 3 till converge.

The number of Gaussian mixture components can be found by using unsupervised clustering algorithm, such as, K - means algorithm. A good initialization estimate for the covariances matrices can be the within-cluster covariances, and a good initialization estimate for the weights can be the fractions of data points allocated to each cluster.

Chapter B

ASV Spoof 2017 Challenge: Baseline System

B.1 Introduction

The CQCC feature set along with the GMM classifier was used as the baseline system for the ASV spoof 2017 challenge. This appendix explains the CQCC feature set in detail. The CQCC feature uses the Constant Q Transform (CQT) for feature extraction. The next section describe the motivation for CQT, the CQT, CQCC feature extraction and results of CQCC+GMM system on ASV spoof 2017 challenge database.

B.2 Motivation for Constant Q Transform (CQT)

The Heisenberg uncertainty principle states that the time and frequency of a signal cannot be computed accurately at the same time (i.e., we cannot achieve best time and frequency resolution simultaneously) [73]. The digital audio signal processing applications use various kinds of time-frequency representations, such as Short-Time Fourier Transform (STFT), Discrete Wavelet Transform (DWT), etc. Among them, the STFT representation is most popular, in which the Fourier transform is applied on short segments of the longer duration signal. The STFT can be interpreted as a filterbank of fixed bandwidth and hence, the Q factor increases as we move from low-to-high frequencies [74]. However, the human perception system has constant Q factor for the frequency range of 500 Hz-20 kHz [75]. Hence, the STFT time-frequency representation is not proper to analyze speech signals. Over the last few decades, the CQT is used for time-frequency representation, which is more similar to human perception system [76]. The CQT resembles with the wavelet transform with high Q factor compared to the wavelet transform. However, the computation complexity of wavelet transform is high. Compared to the wavelet transform, the CQT has less computational complexity.

B.3 Constant Q Transform (CQT)

The CQT $S_{CQT}(k,n)$ of a discrete-time signal $s(n)$ is given as follows [43]:

$$S_{CQT}(k,n) = \sum_{i=n-\lfloor L_k/2 \rfloor}^{n+\lfloor L_k/2 \rfloor} s(i) b_k^*(i - n + L_k/2), \tag{B.1}$$

where k is frequency bin index and $k = 1, 2, ..., K$, $b_k(n)$ are complex-valued basis functions, $b_k^*(n)$ represents complex conjugate of $b_k(n)$m L_k represents the variable window length, and $\lfloor \cdot \rfloor$ denotes floor function. The basis functions $b_k(n)$ are defined as [43]:

$$b_k(n) = \frac{n}{CL_k} e^{j\left(\frac{2\pi n f_k}{f_s} + \phi_k\right)}, \tag{B.2}$$

where f_s is sampling frequency, f_k is center frequency of bin k, ϕ_k is phase offset, and C is scaling factor. For window $w(n)$, the scaling factor C is given as [43]:

$$C = \sum_{j=-\lfloor L_k/2 \rfloor}^{\lfloor L_k/2 \rfloor} w\left(\frac{j + L_k/2}{L_k}\right). \tag{B.3}$$

The center frequency f_k for bin k is given by:

$$f_k = f_1 2^{\frac{k-1}{N_b}}, \tag{B.4}$$

where f_1 is center frequency of lowest frequency bin and N_b is number of bins per octave.

B.4 Constant Q Cepstral Coefficients (CQCC)

Figure B.1 shows the functional block diagram of CQCC feature extraction.

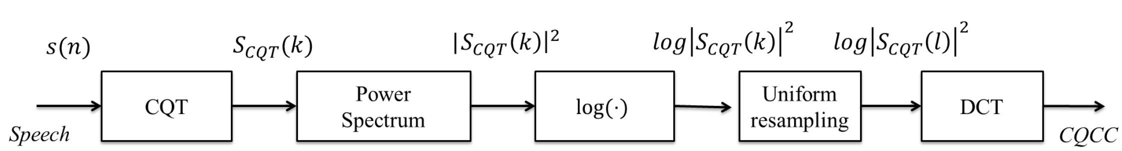


Figure B.1: Functional block diagram of CQCC feature extraction.

The first step of CQCC extraction is CQT. The CQT is applied on the speech signal $s(n)$ to obtain $S_{CQT}(k)$. Furthermore, the power spectrum $|S_{CQT}(k)|^2$ is computed and logarithmic nonlinearity is applied to obtain log power spectrum $log\left(|S_{CQT}(k)|^2\right)$. In the next step, the uniform resampling is performed using polyphase antialiasing filter and a spline interpolation method. Finally, to exploit the redundancy the DCT is applied. Mathematically, CQCC feature extraction process can be written as follows:

$$CQCC(l) = \sum_{m=1}^{M} log\left(|S_{CQT}(k)|^2\right) cos\left[\frac{l(m-1/2)\pi}{M}\right], \tag{B.5}$$

where $l = 0, 1, ..., M-1$ and m are newly sampled frequency bins.

B.5 Experimental Setup and Results

B.5.1 Experimental Setup

The baseline system is built with *90*-D CQCC feature vector that comprise of the zeroth coefficient, 29-static, 30-Δ, and 30-$\Delta\Delta$ coefficients. The minimum frequency set to 15 Hz and maximum frequency to 8 kHz, the number of bins per octave set to 96. On classifier side, GMM uses 512 number of Gaussian components.

B.5.2 Experimental Results

The results of the CQCC+GMM system are shown in Table B.1. The CQCC+GMM system gives EER of 11.89 % and 28.92 % on the development and evaluation set of the ASV spoof 2017 challenge database.

Table B.1: Results for development and evaluation sets using CQCC feature sets

System	EER (%)	
	Development	Evaluation
CQCC	11.89	28.92

List of Publications from the Thesis

1. **Prasad A Tapkir**, Hemant A. Patil, "Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection," *to appear in INTERSPEECH*, Hyderabad, India, Sept. 02-06, 2018.

2. **Prasad A Tapkir**, Ankur T. Patil, Neil Shah, Hemant A. Patil, "Novel Spectral Root Cepstral Features for Replay Spoof Detection," *submitted for possible publication in APSIPA ASC*, Honolulu, Hawaii, USA, 2018.

3. **Prasad A. Tapkir**, and Hemant A. Patil, "Significance of Teager Energy Operator Phase for Replay Spoof Detection," *submitted for possible publication in APSIPA ASC*, Honolulu, Hawaii, USA, 2018.

4. **Prasad A. Tapkir**, Madhu Kamble and Hemant A. Patil, "Replay Spoof Detection uding Power Function based Features," *submitted for possible publication in APSIPA ASC*, Honolulu, Hawaii, USA, 2018.

Other Relevant Publications

1. Nirmesh J. Shah, Neil Shah, Maulik C. Madhavi, Madhu R. Kamble, Hardik B. Sailor, Meet H. Soni, Sreeraj R., **Prasad A. Tapkir**, and Hemant A. Patil, "Novel MMSE-Wasserstein Generative Adversarial Network using Vocal Tract Length Warped Posterior Features for Voice Conversion," *rejected in INTERSPEECH*, Hyderabad, India, Sept. 02-06, 2018.

References

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

[2] P. A. Tapkir and H. A. Patil, "Novel empirical mode decomposition cepstral features for replay spoof detection," accepted in INTERSPEECH, Hyderabad, India, Sept. 02-06, 2018.

[3] P. A. Tapkir, A. T. Patil, N. Shah, and H. A. Patil, "Novel spectral root cepstral features for replay spoof detection," submitted for possible publication in APSIPA ASC, Honolulu, Hawaii, USA, 2018.

[4] P. A. Tapkir and H. A. Patil, "Significance of teager energy operator phase for replay spoof detection," submitted for possible publication in APSIPA ASC, Honolulu, Hawaii, USA, 2018.

[5] 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.

[6] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," National Institute of Standards and Technology (NIST) Gaithersburg MD, Tech. Rep., 1997.

[7] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. Pearson Education, India, 2006.

[8] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Orlando, FL, USA, 2002, pp. IV–4072.

[9] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[10] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[11] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, China, 2004, pp. 145–148.

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

[13] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[14] 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*, New Delhi, India, 2009, pp. 525–530.

[15] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *IEEE Annual Summit and Conference in Asia-Pacific Signal and Information Processing Association (APSIPA-ASC)*, Chiang Mai, Thailand, 2014, pp. 1–5.

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

[17] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 1678–1681.

[18] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *IEEE International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 4, 2011, pp. 1708–1713.

[19] 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.

[20] J. Villalba and Lleida, “Preventing replay attacks on speaker verification systems,” in *IEEE International Carnahan Conference on Security Technology (ICCST)*, Barcelona, Spain, 2011, pp. 1–8.

[21] A. Paul, R. K. Das, R. Sinha, and S. M. Prasanna, “Countermeasure to handle replay attacks in practical speaker verification systems,” in *International Conference on Signal Processing and Communications (SPCOM)*, Bengaluru, India, 2016, pp. 1–5.

[22] R. Font, J. M. Espin, and M. J. Cano, “Experimental analysis of features for replay attack detection - results on the ASV spoof 2017 challenge,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

[23] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, “Novel variable length Teager energy separation based instantaneous frequency features for replay detection,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 12–16.

[24] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, “Spoof detection using source, instantaneous frequency and cepstral features,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 22–26.

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

[26] M. Witkowski, S. Kacprzak, P. Åżelasko, K. Kowalczyk, and J. GaÅĆka, “Audio replay attack detection using high-frequency features,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

[27] X. Wang, Y. Xiao, and X. Zhu, “Feature selection based on CQCCs for automatic speaker verification spoofing,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 32–36.

[28] 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, 2017, pp. 82–86.

[29] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, “Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof 2017,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 87–91.

[30] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 92–96.

[31] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 97–101.

[32] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 102–106.

[33] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

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

[35] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco *et al.*, "REDDOTS replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5395–5399.

[36] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[37] P. P. Vaidyanathan, *Multirate Systems and Filterbanks*, 1st ed. Pearson Education India, 1993.

[38] A. Bouzid and N. Ellouze, "EMD analysis of speech signal in voiced mode," *NOn-LInear Speech Processing (NOLISP)*, p. 112, 2007.

[39] K. Khaldi, A.-O. Boudraa, A. Bouchikhi, M. T.-H. Alouane, and E.-H. S. Diop, "Speech signal noise reduction by EMD," in *Third International Symposium on Communications, Control and Signal Processing (ISCCSP)*. IEEE, 2008, pp. 1155–1158.

[40] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *The Royal Society of London*, vol. 454, 1998, pp. 903–995.

[41] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[42] S. B. 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, pp. 357–366, 1980.

[43] 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, 2016, pp. 249–252.

[44] N. E. Huang, *Hilbert-Huang Transform and Its Applications*. World Scientific, 2014, vol. 16.

[45] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," *The Journal of the Acoustical Society of America (JASA)*, vol. 109, no. 2, pp. 648–670, 2001.

[46] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," in *The Royal Society of London*, vol. 460, 2004, pp. 1597–1611.

[47] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 106, no. 4, pp. 2040–2050, 1999.

[48] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[49] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, vol. 1, Hong Kong, China, 2003, pp. 68–71.

[50] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[51] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 7234–7238.

[52] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[53] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [April 2018] from URL: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, vol. 47, 1997.

[54] 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.

[55] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun 1968.

[56] J. Lim, "Spectral root homomorphic deconvolution system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.

[57] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 556–565, May 2009.

[58] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 106, no. 4, pp. 2040–2050, 1999.

[59] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoustics Research Letters*, vol. 2, no. 3, pp. 91–96, 2001.

[60] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH*, Brighton, United Kingdom (UK), 2009, pp. 28–31.

[61] S. S. Stevens, "On the psychophysical law," *Psychological Review*, vol. 64, no. 3, p. 153, 1957.

[62] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, July 2016.

[63] H. A. Patil and K. K. Parhi, "Development of TEO phase for speaker recognition," in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, Bengaluru, India, 2010, pp. 1–5.

[64] 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.

[65] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear *vs.* mel frequency cepstral coefficients for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, USA, 2011, pp. 559–564.

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

[67] 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.

[68] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, NM, USA, 1990, pp. 381–384.

[69] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[70] S. Mallat, *A Wavelet Tour of Signal Processing*. Second Edition, Academic Press, 1999.

[71] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 617–643, 1992.

[72] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 2007, vol. 382.

[73] 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, 1946.

[74] A. V. Oppenheim, *Discrete-Time Signal Processing*, 2nd ed. Pearson Education India, 1999.

[75] B. C. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[76] 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.