Exploration of Source and Filter Information for the Detection of Replay Attacks in Speaker Verification

Sarfaraz Jelil

Exploration of Source and Filter Information for the Detection of Replay Attacks in Speaker Verification

A

thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

by

SARFARAZ JELIL

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

JANUARY 2025

Certificate

This is to certify that the thesis entitled "**Exploration of Source and Filter Information for the Detection of Replay Attacks in Speaker Verification**", submitted by **Sarfaraz Jelil**, Roll No. 156102027, a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in our opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Prof. Rohit Sinha
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India
Dated: 27/01/2025
Guwahati.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Data Science and Intelligent Systems
Indian Institute of Information Technology Dharwad
Dharwad - 580 009, Karnataka, India.
Dated: 27/01/2025
Dharwad.

To,

My Parents

(***Abdul Jalil and Sara Begum***)

and

My Wife

(***Barasha Nasim***)

Acknowledgements

This thesis would not have been possible without the help and support of several people in various measures. I take this opportunity to express my sincere acknowledgment to all of them.

I express my deepest and most sincere gratitude to my thesis supervisor Prof. Rohit Sinha for his guidance and constant encouragement. His insightful feedback has helped me very much in improving the quality of my thesis. I greatly admire his attitude towards research, creative thinking, and enthusiasm for work. It is truly a blessing to have a supervisor who cares so much about my work and is always there for me in times of need. Without his encouragement and support, completing this thesis would not have been possible. I shall forever remain indebted to him.

I convey my most profound and genuine gratefulness to my co-supervisor Prof. S. R. Mahadeva Prasanna for providing me with the opportunity to work with him after I completed my engineering and motivating me to take up a career in research. I thank him for guiding and supporting me throughout my Ph.D. I shall forever remain indebted to him for playing a vital role in shaping my career.

I am grateful to Prof. P. K. Bora, the chairman of my doctoral committee, for his support and encouragement throughout my Ph.D. period. I am thankful to the other members of my doctoral committee, Dr. Suresh Sundaram and Prof. Priyankoo Sarmah, for sparing their time to evaluate my work. I would also like to thank all other members of the faculty of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their care and support. I am also very thankful to all the technical, office, security, canteen, and maintenance staff members of the department for their help when required.

I am indebted to my seniors (Dr. Rohan Kumar Das, Dr. Sishir Kalita, Dr. Biswajit Deb Sarma, Dr. O.P. Singh) and my dear friends (Abhishek Dey, Mrinmoy Bhattacharjee, Moakala Tzudir, Prabhakar Eedara, Vineeta Das, Samarjeet Das, Sibasis Sahoo, Debasis Jyotishi) and all other members in Signal Informatics and EMST Laboratories. I am also thankful to all other friends of the department for their help and support.

Last but not least, my deepest gratitude goes to my parents and my beloved wife. Without their love, support, and sacrifice, I wouldn't have been able to complete my Ph.D.

Sarfaraz Jelil

Abstract

Automatic speaker verification (ASV) is defined as the task of accepting or rejecting an identity claim of a speaker based on their speech. ASV systems are prone to different kinds of spoofing attacks where the system is presented with a spoofed speech signal instead of a speech signal from a genuine speaker. These spoofing attacks can be a serious threat to an ASV system as they can increase false acceptance rates and negatively impact the performance of the system. Hence, it becomes essential to detect these spoofing attacks and protect the security of an ASV system. This thesis deals with a specific kind of spoofing attack called the replay attack. A replay attack is performed by secretly recording the speech of a genuine user of an ASV system and playing it back to the system to obtain unauthorized access.

A speech signal is produced as a result of the convolution between an excitation source signal and the vocal tract filter. It is hypothesized that the replay attacks will affect both the source and filter components of the speech signal and hence modeling the impact on these components separately will lead to better characterization of replay attacks. Thus, the core of this work is the separation of a speech signal into its components and the design of novel features to capture the replay information present in both components. The initial focus is on utilizing the source component for replay attack detection. To this end, two pitch-synchronous handcrafted source features are proposed which are motivated by a visual analysis of the source component. The handcrafted features are naively defined to extract information only around the glottal closure instants. Thus, the replay information available in the rest of the source signal cannot be utilized resulting in poor performance. This problem is addressed with the proposal of a transform-based pitch-synchronous source feature which is extracted from the region between two adjacent glottal closure instants in the spectral domain.

Extracting features between two glottal closure instants ensures that all voiced regions of the source signal are used. However, it is found experimentally that the unvoiced regions of the speech signal also contain significant replay information. Pitch-synchronous processing of the source signal implies that features can only be extracted from the voiced regions. Two modifications are proposed to overcome the problem involved in pitch-synchronous processing. The first modification is the use of non-pitch-synchronous processing and the second is the extraction of features in the spectro-temporal domain instead of the spectral domain resulting in the development of a novel feature. Finally, the proposed spectro-temporal source feature is augmented with a novel feature extracted from the filter component of the speech signal. The augmentation of the source feature with the filter feature results in the best replay attack detection performance in this thesis. This validates our hypothesis that decomposing a speech signal into source and filter components and modeling both components separately results in a better capture of replay information than modeling the replay attacks directly from the speech signal.

Keywords: Replay attacks, automatic speaker verification, spoofing attacks, ASVSpoof 2017, ASVSpoof 2019.

Contents

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Figures

List of Tables

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Acronyms

AI	artificial intelligence
AM	amplitude modulation
ASP	average statistical pooling
ASR	automatic speech recognition
ASV	automatic speaker verification
CFCCIF	cochlear filter cepstral coefficients-based instantaneous frequency
CILPR	compressed integrated linear prediction residual
CMC	constant-Q multi-level coefficients
CMVN	cepstral mean-variance normalization
CNN	convolutional neural network
CQ-EST	constant-Q equal subband transform
CQ-OST	constant-Q octave subband transform
CQCC	constant-Q cepstral coefficients
CQSPIC	constant-Q statistics-plus-principal information coefficient
CQT	constant-Q transform
CQTMGD	constant-Q transform-based modified group delay
CRNN	convolutional recurrent neural network
CSF gram	combined source filter gram
CSFCC	combined source filter cepstral coefficients
DC	direct current
DCF	detection cost function
DCT	discrete cosine transform
DET	detection error trade-off

DEGG	differenced electro-glottal graph
DF	DeepFake
DFT	discrete Fourier transform
DNN	deep neural network
eCQCC	extended constant-Q cepstral coefficients
EER	equal error rate
EF	epoch feature
EM	expectation maximization
EMST	electro-medical and speech technology
ESA	energy separation algorithm
EU	European union
FAR	false acceptance rate
FFT	fast Fourier transform
FL	Fujisaki-Ljungqvist
FM	frequency modulation
FRR	false rejection rate
GA	glottal activity
GAR	glottal activity regions
GBDT	gradient boosted decision trees
GCI	glottal closure instant
GD	group delay
GIF	glottal inverse filtering
GLC	Gaussian linear classifier
GMM	Gaussian mixture model
GSV	Gaussian mixture model supervector
HFCC	high-frequency cepstral coefficients
HMM	hidden Markov model
HOS	higher-order statistics
HT	Hilbert transform

IA	instantaneous amplitude
IACC	instantaneous amplitude cepstral coefficients
ICQCC	inverted constant-Q cepstral coefficients
IDFT	inverse discrete Fourier transform
IF	instantaneous frequency
IFCC	instantaneous frequency cosine coefficients
IIR	infinite impulse response
ILPR	integrated linear prediction residual
ILRCC	integrated linear prediction residual cepstral coefficients
IMFCC	inverse Mel-frequency cepstral coefficients
JFA	joint factor analysis
LA	logical access
LCNN	light convolutional neural network
LEAF	learnable audio front-end
LF	Liljencrants–Fant
LFCC	linear frequency cepstral coefficients
LNLR	linear-to-non-linear power ratio
LP	linear prediction
LPC	linear prediction coefficients
LPCC	linear prediction cepstral coefficients
LPR	linear prediction residual
MAP	maximum aposteriori
MFCC	Mel-frequency cepstral coefficients
MFM	max feature map
MGD	modified group delay
MLE	maximum likelihood estimation
MMPS	modified magnitude phase spectrum
NAPS	normalized auto-correlation peak strength
OB	occupied bandwidth

OCTAVE	objective control talker verification
OST	octave subband transform
PA	physical access
PCA	principal component analysis
PLDA	probabilistic linear discriminant analysis
PLP	perceptual linear prediction
PSR	peak-to-side-lobe ratio
PSRMS	peak-to-side-lobe ratio mean and skewness
QESA	quadrature energy separation algorithm
RAD	replay attack detection
ReLU	rectified linear unit
RF	random forest
RFCC	rectangular filter cepstral coefficients
RNN	recurrent neural network
SASV	spoofing aware automatic speaker verification
SCD	spectral centroid deviation
SCF	spectral centroid frequency
SCFC	spectral centroid frequency coefficients
SCMC	subband spectral centroid magnitude coefficients
SKR	skewness to kurtosis ratio
SSFC	subband spectral flux coefficients
STFT	short-time Fourier transform
SVM	support vector machine
TDNN	time delay neural network
t-SNE	t-distributed stochastic neighborhood embedding
TTS	text-to-speech
UBM	universal background model
VAE	variational autoencoder
VC	voice conversion

VCTK	voice cloning toolkit
ZFF	zero frequency filter
ZFR	zero frequency resonator

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

1

Introduction

Contents

In an increasingly digital world, establishing the identity of an individual is critical in ensuring authorized access to resources. Biometric authentication is the process of verifying a person's claimed identity based on their physical or behavioural traits such as fingerprint, face, retina, iris, speech, etc. This mode of authentication is more reliable than password-based systems as biometric traits cannot be lost or forgotten [1]. Speech, being the primary means of communication, is one of the preferred choices for biometric authentication. The area of biometrics that uses speech as the mode of authentication is called speaker recognition. Based on the task, speaker recognition can be divided into two types: speaker identification and speaker verification [2]. In speaker identification, given a speech input, the task is to identify which one of the registered users has spoken it. No claim about the user's identity is made in this case. On the other hand, in speaker verification, a claim about the user's identity is also made along with the spoken input. Automatic speaker verification (ASV) refers to a system that accepts or rejects the identity of a speaker using his/her speech. In this thesis, we focus only on ASV systems.

ASV systems are categorized into two types depending upon the text spoken during the enrolment and testing phases of the system, namely text-dependent and text-independent. In text-independent ASV systems, there is no restriction on the recognition phrase used by the speaker in either of the phases [3]. In contrast to this, in a text-dependent ASV system, the lexical content of the test utterance is the same as that used during enrollment [4]. Irrespective of the type of ASV system, it accepts the input speech inserted through a microphone and verifies the claimed identity. The input speech is processed by a front-end which extracts relevant features. The back-end then verifies the claim by comparing the extracted features with the existing features of the speaker already available in a database. These components of an ASV system are depicted in Figure 1.1.

Tremendous progress has been made in the field of ASV research over the years. Thus, recently ASV systems have been adopted as an essential component of applications for secure access such as phone banking, voice-enabled mobile devices, credit card usage, and forensics [5]. Notwithstanding these advancements, ASV systems are susceptible to spoofing attacks, and it has been observed that their performances are severely degraded when subjected to these attacks [6, 7]. In the following sections, an introduction to spoofing attacks and their different kinds is presented.

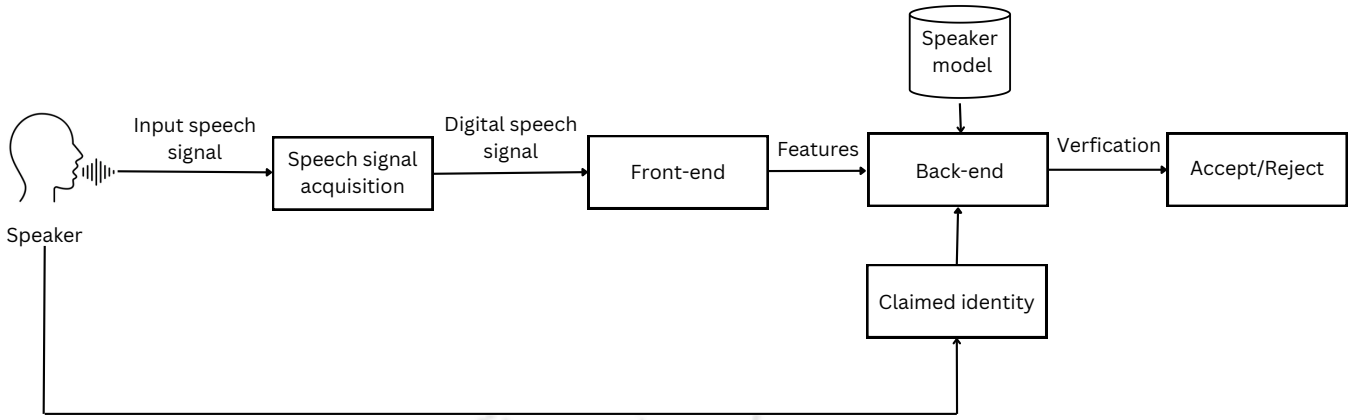

Figure 1.1: Components of an ASV system

1.1 Spoofing Attacks

In the context of an ASV system, a spoofing attack refers to an adversary trying to impersonate an authorized person to bypass and get access to the system [8]. Spoofing attacks can be performed at different points of the ASV system. As shown in Figure 1.2, there are eight possible points of attack. Depending upon the point at which the attack is executed, spoofing attacks can broadly be categorized into two types: direct attacks and indirect attacks [9]. Direct attacks can be performed at the microphone or the transmission level, whereas all other attacks performed at other stages of an ASV system are termed indirect attacks. Since most ASV systems are distributed in nature and implemented via telephony services, it is easy to get access to the microphone and launch a direct attack on the system. Similarly, the speech signal can also be intercepted during the transmission and another signal can be sent to the feature extraction module of the system. However, indirect attacks require system-level access, and hence they are not easy to achieve. Keeping this in mind, the rest of this chapter focuses only on direct attacks.

Considering the modality involved, direct attacks can be classified into four different types, as explained below.

(i) *Impersonation*: It involves an attacker mimicking a person's voice by altering one's voice characteristics to resemble that of the speaker being impersonated.

(ii) *Replay*: In replay attacks, a pre-recorded speech example of the target speaker is played back to the ASV system to gain unauthorized access.

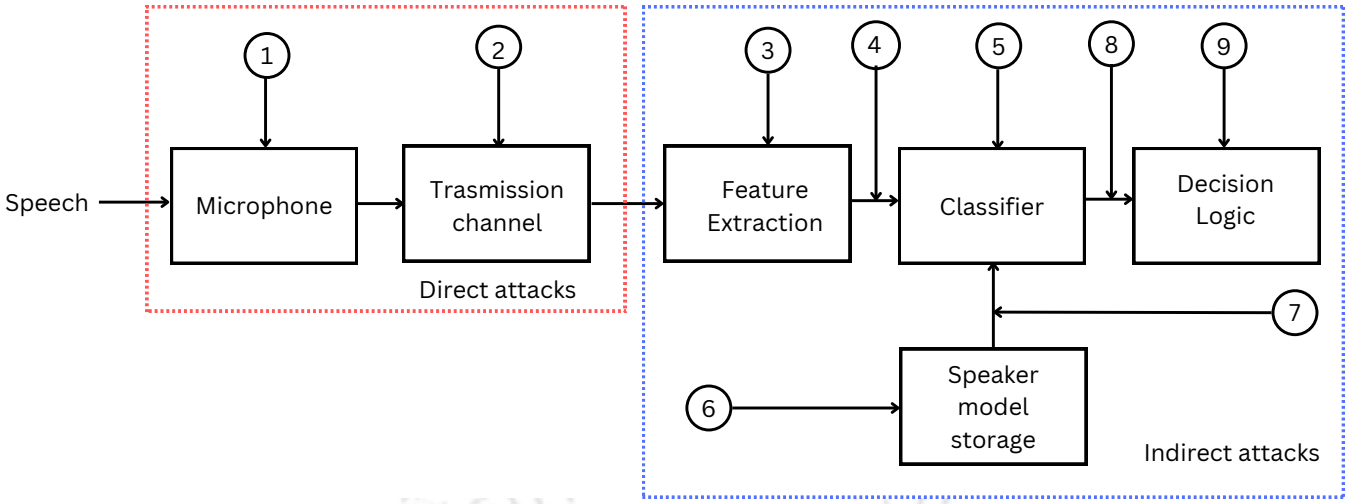

Figure 1.2: Possible points of attacks in an ASV system

(iii) *Speech synthesis*: It is a technique for generating intelligible, natural-sounding artificial speech for any arbitrary text. More commonly, it is referred to as text-to-speech (TTS). The speech generated using the text is used as input to an ASV system in TTS attacks.

(iv) *Voice conversion*: It aims to manipulate the speech of a given speaker so that it resembles, in some sense, that of another target speaker. In contrast to speech synthesis systems which require text input, the input to a voice conversion system is a natural speech signal. Voice conversion (VC) attacks use suitably transformed imposter speech signals to attack an ASV system.

Spoofing attacks can also be classified in terms of the access scenario. The two possible access scenarios include *logical* access (LA) and *physical* access (PA). The TTS and VC attacks are grouped into LA attacks, whereas the replay attacks are grouped into PA attacks. The replay attacks or PA attacks are far easier to perform than synthesizing the speech of any genuine user. They do not require the knowledge of sophisticated speech processing algorithms to generate artificial speech. Moreover, good quality recorders and playback devices are easily and cheaply available. With advancements in the quality of audio recording and playback setups, replay attacks will become more challenging to detect. Hence in this thesis, we deal with the problem of only detecting replay attacks in ASV systems. This problem is referred to as *replay attack detection* (RAD).

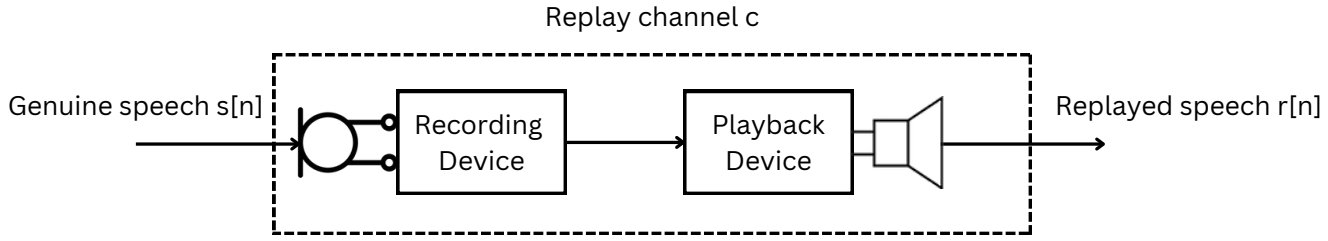

Figure 1.3: Block diagram illustrating a typical replay attack channel

1.2 Characterization of Replay Attacks

In the context of a replay attack on an ASV system, first, the speech of a genuine user is recorded surreptitiously, and then it is played back as input to the system to gain unauthorized access. Therefore, in replay attacks, the genuine speech passes through two channels, namely, the recording channel and the playback channel. Hence, the replayed speech signals are corrupted by the artifacts of both channels. For the sake of modeling, let $c[n]$ denote the composite impulse response of the replay attack channel comprising both recording and playback channels. A replayed signal $r[n]$ can be expressed as a convolution of the genuine speech signal $s[n]$ and the composite impulse response of the replay attack channel $c[n]$. Thus, we can write:

$$r[n] = s[n] * c[n] \tag{1.1}$$

Figure 1.3 shows the block diagram of a typical replay attack channel. The premise behind RAD is the presence of channel noise in the replayed speech signals vis-a-vis the genuine speech signals. For reference purposes, we show genuine and its corresponding replayed speech samples along with their respective spectrograms in Figure 1.4. The primary emphasis in existing works on RAD is to differentiate between the genuine speech signal $s[n]$ and the replayed signal $r[n]$, predominantly using spectral domain features.

It is known that speech is produced by an air stream from the lungs, which goes through the trachea, glottis, and vocal tract. This production mechanism can be mathematically modeled as the convolution between an excitation source signal $e[n]$ generated by the glottal airflow and the impulse response of the vocal tract filter $h[n]$. Thus a genuine speech signal s[n] can be mathematically

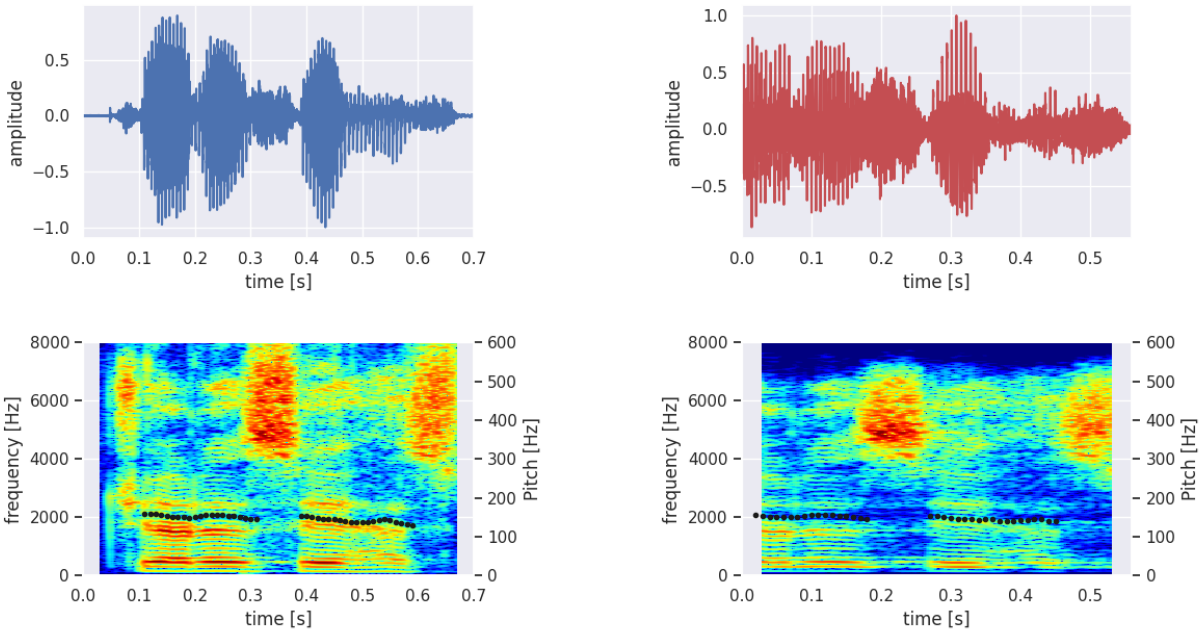

Figure 1.4: Genuine and replayed speech and their corresponding spectrograms along with the pitch contours. It is worth noting that the average pitch values of the genuine and replayed speech signals remain closely similar.

expressed as:

$$s[n] = e[n] * h[n] \tag{1.2}$$

Substituting the value of $s[n]$ from Equation 1.2 into Equation 1.1, we get:

$$r[n] = e[n] * h[n] * c[n] \tag{1.3}$$

We can observe the overall impact of the replay attack channels by analyzing the replayed speech signal. However, to understand how the source and filter components of the speech signal are affected by the replay setup, it would be necessary to decompose the signal. The possible sources of channel noise in replayed signals lie in the distortions introduced in recording and playback channels. In the present times, audio recordings are done digitally rather than analogously, which yields more faithful reproduction. In modern recording and replay devices, the main source of audio noise is the fault or limitation of the involved equipment, which can introduce low-frequency hum, broadband hiss, and nonlinear distortions. Hum is usually caused by electrical interference or when the recording equipment is not properly grounded. Hiss refers to inherent or self-noise present in

the electronic components due to ambient temperature and heat energy generated by the flow of electrons. The nonlinear distortions result from the systems in the audio chain where the output signal is not exactly proportional to the input signal, and harmonics or inter-modulation products are generated. Harmonic distortion covers a wide spectral range.

We hypothesize that the channel noise having differing spectro-temporal signatures would be better captured if the speech signal is divided into source and filter components. More specifically, the harmonic distortions are expected to interact with the vocal filter response, while the hum and/or hiss predominantly affect the excitation component of the replayed speech signal. Motivated by this, the speech signal is decomposed into source and filter components and processed separately to develop different RAD systems.

The decomposition of the speech signal into its constituents is considered vital for many speech-processing tasks. Such a decomposition can be achieved using many different methods, some of which are listed below:

- Pitch-synchronous analysis [10].
- Predictive deconvolution [11].
- Homomorphic deconvolution [12].
- Iterative inverse filter [13].
- System identification-based methods [14, 15].
- Bernoulli Gaussian model-based deconvolution [16].
- Higher order statistics-based inverse filter criteria [17, 18].

Out of these methods, we use the linear prediction (LP) analysis-based decomposition of speech in this thesis to derive novel features for RAD purposes.

1.3 Literature Review on RAD

This section gives a detailed literature review of the countermeasures developed to detect replay attacks. First, we summarize the works done to understand the challenges faced by an ASV system in

the presence of spoofing attacks in Section 1.3.1. Then, a few early attempts to detect replay attacks are discussed in Section 1.3.2. Next, an explanation of the joint efforts of the speaker recognition community to advance research in spoofing detection is provided in Section 1.3.3. These efforts led to the release of different publicly available databases for RAD. Following this, the research done for developing RAD countermeasures using these databases is described. Section 1.3.4 outlines the works done using the ASVSpoof 2017 database. In Section 1.3.5, the efforts made to detect replay attacks utilizing the ASVSpoof 2019 database are detailed. Finally, a few noteworthy contributions for RAD made using the ASVSpoof 2021 database are highlighted in Section 1.3.6.

1.3.1 Vulnerabilities of ASV Systems

- The earliest work on studying the vulnerabilities of an ASV system to replay attacks was done in [19]. To study the impact of replay attacks, a reference ASV system was developed using linear prediction cepstral coefficient (LPCC) features and a hidden Markov model (HMM) classifier. A telephone quality speaker verification database, which consisted of four-digit sequences as passwords, was used in this work. This ASV system is then subjected to replay attacks. The replay speech examples are generated by concatenating isolated digits from the recordings of a genuine user. It was observed that the introduction of replay attacks resulted in a marked increase in both equal error rates (EER) and false acceptance rates (FAR) as compared to those for the ASV system without the presence of replay attacks.

- A similar trend was shown in another study reported in [20]. In this work, the ASV system was developed with mel frequency cepstral coefficient (MFCC) features and a joint factor analysis (JFA) based back-end. Five speakers were used in the database for the experiments. The genuine speech was recorded with a close-talk microphone which was then transmitted over a telephone channel. The replayed signals were recorded simultaneously with the genuine recordings using a far-field microphone. Results from these experiments showed that the EER of the ASV system was 0.71% when only traditional impostor trials were used, but when this EER operating point was selected as the decision threshold, 68% of the replayed trials were falsely accepted by the system. These early studies highlighted the consequences of replay attacks on ASV systems and indicated the need to detect these attacks. Accordingly,

the research into replay attacks gained traction, and several isolated efforts were made to understand and detect replay attacks.

1.3.2 Early Efforts

- One of these efforts includes the work done in [21]. In this work, the basis of detecting replay attacks was the different channel noises present in a replayed speech example in contrast to a genuine speech example. A genuine speech contains the channel noise of the microphone of the ASV system. On the other hand, a replayed speech contains channel noise from the recording and playback devices in addition to the channel noise of the microphone of the ASV system. The channel noise for the genuine and replayed speech is then modeled using a support vector machine (SVM). It was shown that this model was successful in reducing the EER of a Gaussian mixture model - universal background model (GMM-UBM) based ASV system when it was exposed to replay attacks.

- Another approach for detecting replay attacks that use the similarity between a genuine and replayed speech is proposed in [22]. It is based on the hypothesis that a replayed speech for a speaker is recorded from a genuine speech example which is produced during training or testing of the ASV system, the replayed speech will be very similar to one of the stored genuine speech examples. The similarity between the genuine and replayed examples is calculated with the help of spectral bitmaps. A spectral bitmap is extracted from the spectrogram of a speech signal. The spectrogram is divided into non-overlapping blocks, and mean and variance normalization is applied to the amplitude values in each block. Then these values are compared to a predefined threshold and if the value is higher than the threshold it is replaced with a one, otherwise it is replaced with a zero. This new spectrogram consisting of ones and zeros is called the spectral bitmap. To calculate the similarity score, the bitmap of the input speech signal is compared to all the stored bitmaps of the claimed speaker. An element-wise multiplication of the two bitmaps is done, and the sum of this product is considered to be the similarity score. This approach is tested from the perspective of GMM-UBM and HMM-UBM-based ASV systems and was found to be an effective RAD strategy.

- A similar technique using spectral bitmaps in the context of text-independent speaker verifica-

tion was proposed in [23]. Here, instead of having a spectral bitmap for every stored example of a speaker, two averaged spectral bitmaps are generated for the genuine and replayed speech. The spectral bitmap computed from a test speech signal is compared with the two reference spectral bitmaps using the cosine similarity measure. It was found that the proposed method resulted in a high accuracy in detecting replay attacks.

- In the work done in [24], spectral features are utilized to design a replay attack detection algorithm that is robust to adverse acoustic environments. From the spectrogram, five spectral peaks are computed for each frame, and then four peak pairs are generated, considering the highest peak and one other peak at a time. A matrix is constructed where each row represents one pair of peaks. Each row consists of four columns which contain the coordinates of the seed peak and the frequency and time shifts between peaks paired with the seed peak. The matrices of the train and test utterances are then compared to generate a similarity score. This score is further normalized using different normalization techniques to achieve noise robustness.

- A concerted effort to further replay attack detection research began with the organization of a spoofing competition by the biometric group at the IDIAP Research Institute [25]. The competition used the publicly available AVSpoof (audio-visual spoof) database, which contains replay attacks as well as VC and TTS attacks [26]. The organizers also provided a baseline system. The speech signal is short-time processed to derive the power spectrum representation. The computed power spectrum is multiplied with a 40-channel mel filter bank to derive the mel-warped power spectrum. Next, the bands are divided into 10 sub-bands, and the average values for each sub-band are computed in both dimensions. Then, the ratio of these values results in a feature vector which is used to train a logistic regression classifier.

1.3.3 ASVSpoof Challenges

The next significant step taken by the speaker verification community in addressing the challenges involved in spoofing included the organization of a special session at Interspeech 2013 entitled "Spoofing and Countermeasures for Automatic Speaker Verification" [27]. This session brought to attention the need to have common metrics and databases for detecting spoofing attacks so that the different methodologies can be benchmarked against each other. This led to the inception of

ASVSpoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge. The first edition of this challenge was held in 2015 and was called the ASVSpoof 2015 [28]. A database was released as a part of this challenge which included the TTS and VC attacks in a text-independent setting. A new feature called the constant-Q cepstral coefficients (CQCC) was proposed for spoof detection on the ASVSpoof 2015 database in [29]. It is based on the constant-Q transform (CQT) and traditional cepstral analysis. It provides variable time-frequency resolution of speech and hence captures information that other classical features fail to extract. CQCC features, along with a GMM back-end, provided the then state-of-the-art RAD performance.

The second edition, ASVSpoof 2017, focused on detecting only replay attacks on text-dependent ASV systems [30]. A baseline system developed using the CQCC front-end and GMM back-end was released by the organizers. The results from this challenge showed that detecting replay attacks is more difficult than VC and TTS attacks.

The third edition of the challenge, called ASVSpoof 2019, divided the spoofing attacks into logical and physical access scenarios [31]. The logical access included TTS and VC attacks. The physical access scenario was made up of replay attacks. For this edition of the challenges, two baselines based on a GMM classifier and either CQCC or linear frequency cepstral coefficients (LFCC) were adopted.

The most recent edition, ASVSpoof 2021, added DeepFake (DF) attacks in addition to the logical and physical access attacks of the previous challenge [32]. The DF task is introduced to include spoofing in non-ASV system settings. It is designed to emulate the case of an attacker using publicly available speech data of a person and using it to generate speech resembling the voice of that person. The generated speech can contain spoken text which can be used to blackmail the person or harm his/her reputation. For this challenge, four different baselines were used. The first two baselines were the same as that of ASVSpoof 2019. These baselines are RAD systems based on CQCC-GMM and LFCC-GMM. The next baseline used the LFCC features and a light convolutional neural network (LCNN). The final baseline is based on a RawNet2 architecture. It is an end-to-end system and works directly on the raw speech waveforms. The results of the ASVSpoof 2021 challenge showed that the PA and DF tasks are more challenging compared to the LA task.

1.3.4 Survey of Works Using ASVSpoof 2017 Database

- The baseline CQCC features have been compared to several other features in different works. The work done in [33] provided an experimental analysis of eight different features for RAD with a GMM back-end and contrasted these to the CQCC features. The features used in that work are LFCC, MFCC, inverted MFCC (IMFCC), LPCC, rectangular filter cepstral coefficients (RFCC), subband spectral centroid frequency coefficients (SCFC), subband spectral flux coefficients (SSFC) and subband spectral centroid magnitude coefficients (SCMC). These features were evaluated on both the ASVSpoof 2017 and the AVSpoof databases.

- In another work, it was hypothesized that the replay artifacts impact the non-speech region more than the speech region, especially in the high-frequency components of the spectrum [34]. Hence, a feature called the high-frequency cepstral coefficients (HFCC) is proposed. To extract this feature, the speech signal is passed through a second-order high-pass Butterworth filter with a suitable cutoff. The filtered signal is then processed with the classical cepstral analysis to obtain the HFCC features. This feature is used in tandem with CQCC features which are modeled using a deep neural network (DNN) classifier. The DNN is trained in two different ways. In the first method, the DNN has two output nodes corresponding to the genuine and replayed classes. In the second method, the number of output nodes equals the number of replay configurations present in the training data. The authors reported that the second method may have captured more detailed channel information than the binary classifier.

- An ensemble learning technique is proposed in [35]. It uses a combination of acoustic features and different GMM-based classifiers. The CQCC features are post-processed with a mean and variance normalization and then principal component analysis (PCA) is applied to the normalized features. The post-processed CQCC are added to the set of features in addition to the base CQCC features, MFCC, and perceptual linear prediction (PLP) features. The ensemble classifier comprises the GMM, GMM-UBM, ivector-Gaussian PLDA, and GMM supervector-SVM (GSV-SVM) classifiers. In addition to these, two new proposed classifiers are called the GSV-gradient boosted decision trees (GSV-GBDT) and GSV-random forest (GSV-RF). All six classifiers are combined at the score level, and the combined system is

noted to outperform the baseline CQCC-GMM system.

- In another work [36], the authors have explored the efficacy of deep learning frameworks for RAD. Using either FFT or CQT, the normalized log power magnitude spectrum is obtained. These spectrograms are used as input to an LCNN back-end. Another deep learning framework based on the combination of convolutional neural network (CNN) and recurrent neural network (RNN) is also explored in this work. The FFT spectrograms are fed as input into the network. The CNN acts as a feature extractor, and the RNN serves as the back end, which models long-term dependencies.

- Frequency modulation (FM) and amplitude modulation (AM) based features have also been utilized for RAD. In [37], spectral centroid-based frequency modulation features called spectral centroid deviation (SCD) were proposed. In addition to SCD, spectral centroid frequency (SCF) and spectral centroid magnitude coefficient (SCMC) features were also explored in that work. Results showed that the SCD-based RAD system gives more than 60% relative improvement in the EER over a baseline CQCC-GMM system.

- Another approach to detect replay attacks that uses deep Siamese network architecture was examined in [38]. They trained a deep Siamese network to identify pairs of genuine speech signals and pairs of replayed speech signals as similar and pairs of genuine and replayed speech signals as dissimilar.

- In [39], an end-to-end replay attack detection system using deep CNN was proposed. It used a visual attention mechanism on the time-frequency representation of speech signals extracted using group delay features.

- Human cochlear models have also been utilized for the task of RAD. In [40], the authors have proposed cochlear cepstral features derived from energy separation-based instantaneous frequency estimation with a GMM classifier.

- An adaptive-Q cochlear model from which amplitude modulation features were extracted and GMMs were trained for RAD was proposed in [41].

- In [42], a technique using attention-based adaptive filters was proposed that automatically selected only those regions of speech that contained the most discriminative RAD information.

1.3.5 Survey of Works Using ASVSpoof 2019 Database

- Long-range acoustic features derived using long-term CQT were used for RAD in [43]. These features are the extended CQCC (eCQCC), inverted CQCC (ICQCC), constant-Q multi-level coefficient (CMC), constant-Q equal subband transform (CQ-EST), constant-Q octave subband transform (CQ-OST) and constant-Q statistics-plus-principal information coefficient (CQSPIC). The eCQCC feature is a combination of CQCC features computed from the linear power spectrum and the coefficients calculated from the octave power spectrum. On the other hand, the ICQCC features are extracted from the inverse of the long-term CQT linear power spectrum. The CQ-EST and CQ-OST are subband features, and the CQSPIC combines extended coefficients of eCQCC and CQ-OST features coupled with the variance. These features are used to develop RAD systems using either a GMM or a DNN classifier.

- Features extracted from the time-frequency representation of speech gained prominence for developing RAD systems on the ASVSpoof 2019 database. In [44], the log-power spectrum is calculated from the speech signal using the CQT, FFT, and discrete cosine transform (DCT). These features are used as input to an LCNN classifier to detect replay attacks.

- Magnitude and phase-based time-frequency representations were used to extract features in [45]. The magnitude-based features included the log-power spectrum extracted from FFT, mel scale filter banks, and the CQT-based log-power spectrum. The phase-based features included the modified group delay features (MGD) and the proposed CQT-based MGD (CQT-MGD). These features were used with a ResNeWt architecture for RAD.

- A similar set of features were used in [46]. Since a speech signal is characterized completely using both the magnitude and phase spectrum, a joint feature modeling of both these components is proposed. The short-time FFT-based gram features and the group delay (GD) gram are combined to obtain the joint feature.

- In [47], different spectral features, including MFCC, IMFCC, CQCC, and SCMC, are utilized

along with several shallow and deep classifiers such as GMM, CNN, SVM, convolutional recurrent neural network (CRNN), and Wave-U-Net are utilized. The authors propose using an ensemble of these models to achieve the best performance.

- AM and FM-based features have also been proposed for RAD in [48]. These features are extracted from the IA and IF components of filtered subband speech signals with the help of the energy separation algorithm (ESA) and are called the instantaneous amplitude cepstral coefficients (ESA-IACC) and instantaneous frequency cepstral coefficients (ESA-IFCC). These features are coupled with a GMM classifier to develop different RAD systems.

- In another work, a novel feature called the cochlear filter cepstral coefficients-based instantaneous frequency using quadrature energy separation algorithm (CFCCIF-QESA) was proposed [49]. It is extracted by using relative phase shift to compute the IF and provides excellent temporal resolution and relative phase information. This feature is used to build RAD systems using GMM, CNN, and LCNN classifiers.

- Multi-task learning has also been utilized for RAD. The authors of [50] have used it to study its effect on the generalizability and discriminability of RAD systems. They optimized a residual network (ResNet) with multi-task learning using Siamese neural networks.

- A combination of phase and magnitude spectra information for RAD was proposed in [51]. A feature named modified magnitude-phase spectrum was proposed that can model both phase and spectrum from the speech signal.

- In [52], two back-ends are proposed that use the output probabilities (scores) from the GMM trained from the genuine and replayed speech as input. The models are called GMM-ResNet and GMM-Squeeze excitation network (GMM-SENet). They model the relationship between the frames in addition to modeling the score distribution.

1.3.6 Survey of Works Using ASVSpoof 2021 Database

The PA training and development data for the ASVSpoof 2021 challenge is the same as that of the ASVSpoof 2019 dataset. The PA data for the ASVSpoof 2019 database are collected in

simulated replay environments. However, the evaluation data for ASVSpoof 2021 are collected in real replay environments. Hence, there is a mismatch between the training and evaluation data in the ASVSpoof 2021 challenge, and the main research issue is to develop RAD systems that are robust to this environmental mismatch.

- To deal with the differences in the replay channels in the training and evaluation data, one-class classification is used in [53]. The data from only the genuine class is used to train the models, which enhances their generalization capability against unseen attacks. Front-end features extracted using a vocoder are used in this work. It is assumed that the vocoder will be able to eliminate the replay channel information without reducing the speaker and text information in the input speech. The vocoder-filtered speech signal is taken as the reference signal. Next, the log spectrogram of the reference signal is subtracted from that of the original speech signal and is used as the input feature. This subtraction will reduce the speaker and text information and highlight the replay channel information. These features are used with one-class GMM and variational autoencoder (VAE) models.

- A time delay neural network (TDNN) is used for RAD in [54]. The network is used along with front-end features like MFCC, CQCC, SCMC, and linear filterbank energies to develop different RAD systems. The embeddings computed from the MFCC-based TDNN are used as input features for different classifiers like GMM, SVM, and Gaussian linear classifier (GLC).

- Mel spectrogram with various numbers of frequency bins and learnable audio front-end (LEAF) is used as the feature extractor in [55]. Classifiers are implemented via ResNet, whose outputs are forwarded to scores via attentive statistical pooling (ASP). The final score is obtained via empirical weight averaging.

- In the work done in [56], the log magnitude spectrogram computed using the short-time Fourier transform is used as a feature. These features are fed into a squeeze excitation residual network (SE-ResNet) inspired from [57]. The networks are trained using a generalized end-to-end loss, and two post-processing methods are introduced, which increases the RAD performance.

- The use of a multiple-point CNN for RAD is proposed in [58]. Acoustic features derived from speech are of variable lengths owing to the difference in the length of the utterance. The

multiple-point CNN is used to overcome the problem of handling features of different sizes when training the CNN [59]. Another proposal in this paper is to use the phase spectra of both the original speech signal and the time-inverted version of the signal [60]. Four types of CNN-based networks are used in this work. These are SE-ResNet [61], DenseNet [62], ShuffleNetV2 [63] and MNASNet [64].

The literature survey revealed that the primary method to detect replay attacks is to design a suitable front-end feature and couple it with an appropriate back-end classifier. Several end-to-end systems have also been proposed for RAD. It was seen that various features were explored for RAD with either a classical back-end (such as GMM, SVM, etc.) or a deep learning-based back-end. Further, the two commonly used databases used for RAD research are the ASVSpoof 2017 and the PA set of the ASVSpoof 2019. Keeping in sync with the above observation, in this thesis, we also attempt to design a few unique features for RAD. These features are used with a GMM and a deep learning-based classifier. Finally, the proposed features and the corresponding back-end are evaluated on the ASVSpoof 2017 and the PA set of the ASVSpoof 2019 database.

1.4 Problem Formulation

From the literature review, we find that most of the works have attempted RAD through direct characterization of the speech. No attempt has been made to decompose the speech so far. Motivated by the hypothesized characterization of replay attacks through the decomposition of speech discussed in Section 1.2, we pursue the following objectives in this thesis.

- To establish that the processing of the source component of speech signals provides a viable means for RAD.
- To design techniques to better characterize the source signals for RAD purposes.
- Studying the relative contribution of voiced and unvoiced regions of the speech for RAD.
- Exploring the joint modeling of the decomposed source and filter components of the speech signal for RAD purposes.

The thesis is built upon the hypothesis that replay artifacts impact the source (excitation) and filter (vocal tract) components of speech differently, motivating the use of source-filter decomposition for replay attack detection. This foundational idea forms the basis for a systematic and logical sequence of studies aimed at addressing this challenge.

The initial study investigates source-based features, beginning with pitch-synchronous approaches. Epoch-based features, extracted using zero frequency filtering, are designed to capture distortions in the excitation signal that occur due to replay artifacts. Recognizing that pitch-synchronous features are limited to voiced regions, the study is extended to non-pitch-synchronous features that analyze the entire speech signal, including unvoiced regions, using spectro-temporal representations of the source signal.

While the source-based features reveal important characteristics of replay artifacts, they are primarily sensitive to distortions in the excitation component of the speech signal. However, replay attacks also introduce significant changes to the vocal tract characteristics, which are not fully captured by source-based features. This realization motivates the exploration of filter-based features, which aim to model distortions introduced in the filter component by replay attacks.

The study of filter-based features uses linear prediction coefficients to model the vocal tract's spectral response, capturing distortions introduced by the nature of the replay setup. These studies of source and filter components establish a comprehensive understanding of replay artifacts and their distinct impact on the two components of speech.

The logical progression leads to the development of combined source-filter features. These features integrate complementary information from the source and filter components, leveraging their distinct sensitivity to replay artifacts.

Finally, the studies culminate in a comprehensive experimental evaluation using benchmark datasets like ASVSpoof 2017 v2.0 and ASVSpoof 2019. These experiments validate the effectiveness of the proposed methods and highlight their robustness across varying acoustic conditions, completing the systematic sequence of studies motivated by the potential of source-filter decomposition.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows. A detailed description of the databases used for the experiments and analysis performed in this work is given in Chapter 2. It also discusses the baseline systems used for contrast purposes in this work. First, an explanation of the baseline features and the Gaussian mixture model (GMM) classifier for RAD is provided. Next, the experimental setups of the baseline RAD systems and their performances on the databases are reported.

In Chapter 3, the initial attempts made to detect replay attacks are presented. The differences in the nature of the genuine and replayed source signals are studied first with the help of source characteristics. Next, two handcrafted features for RAD based on the traditional pitch-synchronous processing of the source signal are proposed in this chapter. Using these two features and the GMM classifier, two RAD systems are developed on the ASVSpoof 2017 v2.0 database. A comparison of the performance of the proposed systems with the baseline systems is then provided.

Chapter 4 explores another representation of the source signal called the integrated linear prediction residual (ILPR). An explanation of how the glottal source dynamics can be captured from this signal is given. Next, a pitch-synchronous transform-based source feature for RAD is proposed. This feature is computed from the ILPR signal between two glottal closure instants. A RAD system is developed with this feature and a GMM back-end on the ASVSpoof 2017 v2.0 database and the results of the experiments are then presented. Following that, an ablation study is done to show the relevance of voiced and unvoiced regions of speech for RAD.

A non-pitch-synchronous processing of the source signal is introduced in Chapter 5. First, a detailed analysis is presented to illustrate the effect of processing the source signal non-pitch-synchronously. Following that, a feature derived from the spectro-temporal representation of the source signal is proposed. This feature allows the use of both voiced and unvoiced regions of speech, unlike the previously used features. RAD systems are developed using this feature and a GMM back-end on both the ASVSpoof 2017 v2.0 and ASVSpoof 2019 databases.

The hypothesis that the decomposition of the speech signal into source and filter components gives enhanced RAD performance is tested in Chapter 6. A feature for the filter component is first proposed and evaluated on the ASVSpoof 2017 v2.0 database. Next, this feature is combined with the non-pitch-synchronous source feature described earlier. A RAD system is developed with the

combined feature and a GMM back-end on the ASVSpoof 2017 v2.0 database,, and the results are reported. Following this, a ResNet-18-based back-end is used to develop a RAD system using the combined feature. This system is evaluated on the ASVSpoof 2019 database. Then, the source and filter features are tweaked to obtain source-filter time-frequency features. The source and filter features are also combined and a RAD system is developed with a ResNet18 back-end. The performance of this RAD system on the ASVSpoof 2019 database is reported.

Finally, the thesis is summarized, and the possible directions of future work are discussed in Chapter 7.

2

Databases and Baseline RAD Systems

Contents

2.1 Introduction

In Chapter 1, we provided an introduction to replay attacks and how an ASV system can be impacted by such attacks. It is essential to build a RAD system to deal with these attacks. Hence, this thesis focuses on the development of RAD systems by trying to decompose the speech signal and exploiting the replay information present in both the source and filter components of the signal. The relevant source and filter information is extracted from the speech signal using several proposed features. It is necessary to have data of good quality and quantity to develop the RAD systems using the proposed features and evaluate the efficacy of these features. Thus, in Chapter 2, we describe the two replay attack databases used in this thesis to develop and evaluate the RAD systems built with the proposed features. The two databases are the ASVSpoof 2017 v2.0 [65] and ASVSpoof 2019 [66]and have been collected as part of the ASVSpoof challenges.

To understand the performance improvement obtained by the RAD systems developed using the proposed features, it is necessary to have a baseline RAD system. There are two baseline systems that are used in this thesis and they are discussed in the chapter. The baseline systems are developed using either a CQCC or LFCC front-end with a GMM-based back-end. The performances of the baseline systems in terms of EER and t-DCF are also provided.

The remainder of the chapter is organized as follows: Section 2.2 describes the ASVSpoof 2017 v2.0 database in detail. In Section 2.3, the particulars of the ASVSpoof 2019 database are presented. An explanation of the CQCC and LFCC features and the GMM is given in Section 2.4. The experimental setup of the baseline systems is described in Section 2.5. In Section 2.6, the results of the baseline systems are reported. Finally, the conclusions drawn from this chapter are provided in Section 2.7.

2.2 ASVSpoof 2017 Database

The ASVSpoof 2017 database was designed to support the development of countermeasures to protect ASV systems from replay spoofing attacks. This database is an evolution from its predecessor, ASVSpoof 2015, incorporating a more extensive collection process and a broader range of replay scenarios to simulate real-world conditions better.

2.2.1 Data Collection

The ASVSpoof 2017 database comprises bona fide and spoofed utterances [30]. Bona fide utterances are sourced from the RedDots corpus, which contains recordings from volunteers using Android smartphones [67]. These recordings feature one of ten different fixed pass-phrases, providing a consistent basis for comparison.

Spoofed utterances are created by replaying and recording these bona fide utterances using various devices and in different acoustic environments. This process is intended to simulate realistic replay attacks, capturing the diversity in playback and recording equipment and environments that might be encountered in real-life scenarios. The replay utterances were collected by multiple participants, with 57% contributed by members of the EU Horizon 2020-funded OCTAVE (objective control for talker verification) project and the remaining 43% by other collaborators.

2.2.2 Data Partitions

The database is divided into three main subsets: training, development, and evaluation. Each subset contains both bona fide and spoofed recordings, but they differ in sizes and compositions. The partitions are designed to facilitate different stages of system development and testing.

- Training set: Contains a smaller, controlled set of recordings intended for initial system training and tuning.
- Development set: Used for validation during system development, this subset provides a slightly larger and more varied collection of recordings.
- Evaluation set: The largest and most diverse subset, used for the final evaluation of system performance.

Meta-data including ground-truth labels (bona fide/spoofed), speaker IDs, phrase IDs, and replay configuration details are provided for the training and development subsets. For the evaluation subset, these details were initially withheld to ensure unbiased assessment but were later released with version 2.0 of the database to support more detailed performance analysis.

Table 2.1: Details of the ASVSpoof 2017 v2.0 database.

Database Subset	No. of Speakers	No. of Utterances	
		Genuine	Spoofed
Train	10	1507	1507
Development	8	760	950
Evaluation	24	1298	12992

2.2.3 Database Update

The original ASVSpoof 2017 database [30] was later updated to address several data anomalies identified post-challenge evaluation. The updated database is known as the ASVSpoof 2017 v2.0 database [65]. These updates were necessary to ensure the integrity of the dataset and the validity of experimental results. The anomalies included issues such as inconsistencies in recording conditions and errors in the labeling of replay configurations.

Version 2.0 provides a corrected and more reliable dataset, enabling researchers to conduct more accurate evaluations of their spoofing countermeasures. This version also includes enhanced metadata, allowing for a deeper analysis of system performance across different acoustic environments and device configurations.

Table 2.1 summarizes the composition and details of the ASVSpoof 2017 Version 2.0 database:

2.2.4 Data Collection Process

The data collection process for ASVSpoof 2017 was meticulously planned to ensure a representative and diverse dataset. Here are the key steps involved:

- Selection of Bona Fide Utterances: The RedDots corpus provided a solid foundation of bona fide recordings, ensuring consistency in the content of utterances (fixed pass-phrases) and capturing variations in speaker characteristics [67].

- Replay Attack Simulation: To generate spoofed utterances, bona fide recordings were played back using different devices (e.g., smartphones, laptops) and recorded in various acoustic environments. This simulation aimed to cover a wide range of potential replay attack scenarios, including different room sizes, background noise levels, and device qualities.

- Data Annotation and Verification: Each recording was carefully labeled with ground-truth information, specifying whether it was bona fide or spoofed. Additional meta-data such as speaker ID, phrase ID, and replay configuration details were recorded to facilitate detailed analysis and comparison.

2.2.5 Metadata

The comprehensive meta-data provided with the ASVSpoof 2017 database plays a crucial role in supporting research and development of spoofing countermeasures. The meta-data includes:

- Ground-Truth Labels: Indicating whether an utterance is bona fide or spoofed.
- Speaker IDs: Unique identifiers for each speaker in the database.
- Phrase IDs: Identifiers for the fixed pass-phrases used in the recordings.
- Replay Configurations: Information about the devices and acoustic environments used for simulating replay attacks.

This detailed meta-data enables researchers to perform in-depth analyses, such as evaluating the performance of countermeasures in different acoustic settings or with different types of replay devices.

2.2.6 Applications and Impact

The ASVSpoof 2017 database has had a significant impact on the field of ASV research, particularly in the development of countermeasures against replay attacks. By providing a realistic and challenging dataset, it has enabled researchers to test and refine their systems for conditions that closely mimic real-world scenarios.

Several key applications and areas of impact include:

- Benchmarking and Evaluation: The database serves as a benchmark for evaluating the performance of ASV systems and their robustness against spoofing attacks. Researchers can compare their methods using a common dataset, facilitating the advancement of the field.

- Development of Countermeasures: The diversity and complexity of the ASVSpoof 2017 database have driven innovation in countermeasure techniques, leading to more sophisticated and effective methods for detecting replay attacks.

- Understanding Replay Attacks: The detailed meta-data and varied replay configurations help researchers understand the factors that influence the success of replay attacks, providing insights into how to design more resilient ASV systems.

2.2.7 Performance Metrics

The primary metric used to compute the performance of RAD systems using the ASVSpoof 2017 v2.0 database is the equal error rate (EER). The EER is computed in the following way.

Let $P_{fa}(\theta)$ and $P_{miss}(\theta)$ be the false alarm and miss rates at threshold θ defined according to:

$$P_{fa}(\theta) = \frac{\text{No. of replay trials with score} > \theta}{\text{Total no. of replay trials}} \tag{2.1}$$

$$P_{miss}(\theta) = \frac{\text{No. of non-replay trials with score} \leq \theta}{\text{Total no. of non-replay trials}} \tag{2.2}$$

where $P_{fa}(\theta)$ and $P_{miss}(\theta)$ are the monotonically decreasing and increasing functions of θ, respectively. The EER corresponds to the threshold θ_{EER} at which the two detection error rates are (approximately) equal.

2.3 ASVSpoof 2019 Database

The ASVSpoof 2019 database, the third in its series, introduces a comprehensive dataset designed to tackle three major spoofing techniques: replay, speech synthesis, and voice conversion [66]. The replay attack corresponds to the physical access scenario whereas the speech synthesis and voice conversion attack corresponds to the logical access (LA) scenario. Since this thesis focuses only on replay attacks, we provide an in-depth description of the physical access (PA) part of the ASVSpoof 2019 database.

2.3.1 Physical Access Scenario

In contrast to logical access, the physical access scenario deals with replay attacks where previously recorded bona fide speech is played back to the ASV system. These attacks are presented through the acoustic environment, introducing variabilities such as room acoustics, background noise, and recording/playback device characteristics. This scenario is highly relevant for real-world applications where ASV systems are deployed to secure physical spaces like bank vaults, secure facilities, and restricted areas.

2.3.2 Database Collection

The PA part of the ASVSpoof 2019 database is meticulously designed to simulate a wide range of real-world replay attack conditions. It includes recordings made in different acoustic environments, with variations in the distance between the speaker and the microphone, room sizes, and levels of background noise. This comprehensive approach ensures that the database covers a broad spectrum of potential replay attack scenarios, providing a robust platform for developing and testing spoofing countermeasures.

2.3.2.1 Acoustic Environments

The PA subset was designed to simulate real-world conditions where ASV systems might be deployed, capturing the variability in room acoustics and recording conditions. To this end, 27 different acoustic environments were defined, each characterized by three parameters: room size (S), reverberation level (R), and talker-to-ASV distance (Ds). These parameters were divided into three categories each, resulting in the following combinations:

- Room Size (S): Small (2-5 m^2), Medium (5-10 m^2), Large (>10 m^2)
- Reverberation Level (R): Low (< 200 ms), Medium (200-500 ms), High (> 500 ms)
- Talker-to-ASV Distance (Ds): Close (10-50 cm), Intermediate (50-100 cm), Far (>100 cm)

2.3.2.2 Replay Attack Configurations

Replay attacks were simulated by varying the distance between the attacker and the target speaker (Da) and the quality of the replay device (Q). Each of these parameters was also divided

into three categories:

- Attacker-to-Talker Distance (Da): Close (10-50 cm) Intermediate (50-100 cm) Far (>100 cm)
- Replay Device Quality (Q): High (e.g., professional loudspeakers) Medium (e.g., consumer-grade devices) Low (e.g., low-cost, low-quality devices)

The combination of these factors results in 9 different replay configurations (AA, AB, AC, ... CC), providing a comprehensive assessment of how varying replay conditions affect ASV systems.

2.3.2.3 Recording Procedure

For both bona fide and replay recordings, speech utterances from the voice cloning toolkit (VCTK) corpus were used, sampled at 96 kHz, and downsampled to 16 kHz [68]. The recording setup involved using a fixed microphone and varying the position of the speaker to simulate different talker-to-ASV distances. Replay attacks were created by first recording the original utterances in various acoustic environments and then playing them back using different loudspeaker devices.

2.3.2.4 Acoustic Environment Simulation

The simulation of acoustic environments was performed using Roomsimove, a room acoustics simulator. This tool allowed for precise control over the parameters of each environment (S, R, Ds), ensuring that each combination was accurately represented.

Bona fide recordings were created by simulating the acoustic environment effects without any additional processing. This involved positioning the speaker at various distances from the microphone and recording the speech in different room conditions.

Replay attacks were simulated by first recording the original speech in a controlled environment and then playing it back through different loudspeakers at varying distances from the microphone. This process involved several steps:

- Recording the Original Speech: The original speech utterances were recorded in a quasi-anechoic environment to ensure clarity and high quality.
- Simulating Room Effects: The recorded speech was then processed using Roomsimove to simulate the effect of different room acoustics.

- Playback Through Loudspeakers: The processed speech was played back using various loudspeakers positioned at different distances from the microphone, simulating real-world replay attack scenarios.

2.3.2.5 Replay Device Characteristics

The characteristics of each replay device were meticulously documented. Each device was tested in a controlled environment to measure its occupied bandwidth (OB), minimum frequency (minF), and linear-to-non-linear power ratio (LNLR). These measurements helped categorize the devices into high, medium, and low quality, ensuring a diverse range of replay conditions.

Each replay device was set up in a consistent manner to maintain the integrity of the data. The loudspeakers were placed at predefined distances from the microphone, and the volume levels were standardized to ensure uniformity across all recordings.

2.3.3 Structure of the PA Set

The PA set is divided into three partitions: training, development, and evaluation. Each partition includes both bona fide (genuine) and spoofed (replayed) utterances, allowing researchers to train, tune, and evaluate their ASV systems and countermeasures.

- Training Set: The training set is used to develop and train spoofing countermeasures. It includes a balanced mix of bona fide and spoofed utterances, enabling researchers to understand the characteristics of both types of speech and design features and models that can effectively distinguish between them.

- Development Set: The development set is used for tuning and validating the performance of the developed countermeasures. It provides a separate dataset that helps in fine-tuning the models and adjusting parameters to achieve optimal performance in detecting replay attacks.

- Evaluation Set: The evaluation set is used for the final assessment of the ASV systems and countermeasures. It includes a diverse range of replay attack scenarios, ensuring that the performance metrics reflect the system's robustness against a wide variety of real-world conditions.

Table 2.2 summarizes the salient details of the PA partition of the ASVSpoof 2019 database:

Table 2.2: Table showing the details about the number of speakers and the number of utterances for the PA part of ASVSpoof 2019 database

Portion	Subset	No. of Speakers		No. of Utterances	
		Male	Female	Bonafide	Spoof
ASVSpoof 2019 PA	Train	8	12	5,400	48,600
	Development	8	12	5,400	24,300
	Evaluation	30	37	18,089	134,630

2.3.4 Impact on ASV Systems

The physical access scenario presents significant challenges to ASV systems due to the added complexity of the acoustic environment. The variability in room acoustics, background noise, and device characteristics can significantly affect the system's ability to accurately distinguish between bona fide and replayed speech. The PA part of the ASVSpoof 2019 database provides a comprehensive framework for testing and improving the resilience of ASV systems against these challenges.

2.3.5 Performance Metrics

The primary performance metrics used in the ASVSpoof 2019 challenge are the equal error rate (EER) and the tandem detection cost function (t-DCF). The EER measures the point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal, providing a single metric for overall system performance. The t-DCF metric, on the other hand, evaluates the combined performance of the ASV system and the replay countermeasures, reflecting their joint effectiveness in mitigating replay attacks. A detailed explanation of the t-DCF metric can be found in [69].

2.4 Baseline Features and Back-end

This section describes the process of extracting the CQCC and LFCC features from a speech signal. Next, the GMM is explained with the relevant mathematical equations.

2.4.1 Constant-Q Cepstral Coefficients

In speech signal analysis, the use of an optimal time-frequency representation is central. The fundamental premise lies within the principles of the uncertainty principle, as delineated by Equation 2.3, which posits that the precision in measuring time and frequency content is inherently

limited [70]. This principle underscores a critical trade-off: as frequency resolution (Δf) increases, temporal resolution (Δt) decreases, and vice versa, where the product of these resolutions remains constant. Thus, the temporal and frequency characteristics of a signal are inextricably intertwined, requiring a careful selection of representation for meaningful analysis.

$$\Delta f \Delta t \geq \frac{1}{4\pi} \tag{2.3}$$

One of the most widely used tools in digital signal processing to obtain the time-frequency representation is the short-time Fourier transform (STFT). The STFT processes a speech signal by dividing it into shorter segments via a sliding window, thereby calculating its local frequency content over time. The Q factor is a measure of the selectivity of each filter and is calculated as the ratio between the centre frequency f_k and the bandwidth δf.

$$Q = \frac{f_k}{\delta f} \tag{2.4}$$

However, the STFT has a fixed bandwidth per filter and yields an increasing Q factor with ascending frequencies. It is also known that the human perception system approximates a constant-Q factor between 500 Hz and 20 kHz [71]. Thus, the STFT diverges from the constant-Q factor of the human perception system within the audible range.

The problem of increasing Q factor can be solved by using the constant Q-transform (CQT), a perceptually motivated alternative that offers a more congruous framework for speech signal analysis [72]. Unlike the STFT, which exhibits a fixed time-frequency resolution, the CQT dynamically adjusts its frequency bins to adhere to the perceptual uniformity of human hearing. The CQT capitalizes on geometrically distributed octaves and center frequencies, mirroring the frequency perception continuum. The efficiency of the CQT was further augmented by aligning center frequencies with the equal-tempered scale of western music, enhancing its utility in music signal processing applications [73].

The salient advantage of the CQT lies in its adaptive resolution, offering higher frequency resolution for lower frequencies and higher temporal resolution for higher frequencies. This adaptability, coupled with relatively high Q factors (about 100 bins per octave), renders the CQT a potent tool for the analysis, classification, and separation of audio signals, as evidenced by its wide-ranging

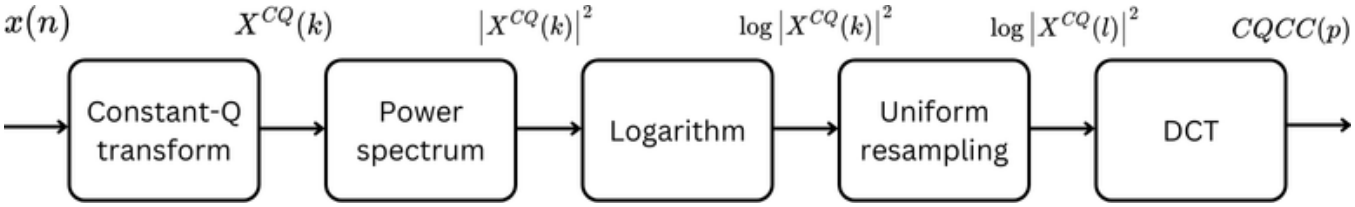

Figure 2.1: Components of an ASV system

applications in recent years.

The computation of CQCC features from the CQT of a speech signal is discussed next. Let $s_m(n)$ denote the m^{th} frame of the input speech and $X_m(k)$ the corresponding constant-Q spectrum which is computed as [29]:

$$X_m(k) = \sum_{n=m-\lfloor N_k/2 \rfloor}^{m+\lfloor N_k/2 \rfloor} x(n) a_k^*(n - m + N_k/2) \tag{2.5}$$

where, $0 \leq k \leq K-1$ is the frequency bin index, a_k^* represents the complex conjugate of a_k and N_k denotes variable window lengths.

The resulting CQT spectrum is then converted to a logarithmically compressed CQT power spectrum which is then uniformly sampled to obtain a logarithmic compressed linear power spectrum as [74].

$$\log|X_m(k)|^2 \longrightarrow \log|X_m(l)|^2 \tag{2.6}$$

where $0 \leq l \leq L-1$ are newly re-sampled linear frequency bins. Applying DCT to the logarithmic linear power spectrum yields the CQCC feature $Y_m(p)$ as

$$Y_m(p) = \sum_{l=0}^{L-1} \log|X_m(l)|^2 \cos\left[\frac{\pi(2l+1)p}{2L}\right] \tag{2.7}$$

where $0 \leq p \leq L-1$.

The block diagram of the process of extraction of CQCC features from a speech signal is given in Figure 2.1.

2.4.2 Linear Frequency Cepstral Coefficients

In this thesis, another feature called the linear frequency cepstral coefficients (LFCC) is used to develop RAD systems in addition to the CQCC features. LFCC features are similar to the

mel-frequency cepstral coefficients (MFCC) features and both of these have been used for speaker verification. The difference between these two features is in the way they handle the frequency scale. The MFCC features use a mel-scale frequency axis, which is a non-linear scale that mimics the human auditory system's perception of pitch [75]. On the other hand, LFCC uses a linear frequency scale. Unlike MFCC, which groups frequencies nonlinearly, LFCC maintains a linear spacing between frequency bands.

The steps in the extraction of LFCC features from a speech signal are described below:

- Pre-emphasis: The speech signal $s[n]$ is first pre-emphasized to amplify higher frequencies, which helps in improving the signal-to-noise ratio. This is usually done by applying a first-order high-pass filter as shown in the Equation 2.8

$$y[n] = s[n] - \alpha \cdot s[n-1] \tag{2.8}$$

  where, y[n] is the pre-emphasized signal and α is the pre-emphasis coefficient (typically around 0.95).

- Frame blocking: The pre-emphasized speech signal is divided into short overlapping frames. This is typically done using a window function like the Hamming window to reduce spectral leakage. Let $w[n]$ be the window function, then the windowed speech signal $y_w[n]$ is obtained as shown in Equation 2.9.

$$y_w[n] = w[n] \cdot y[n] \tag{2.9}$$

- Discrete Fourier transform: The discrete Fourier transform (DFT) is applied to each frame to convert it from the time domain to the frequency domain. This gives the spectrum $Y(k)$ of the frame as depicted in the Equation 2.10:

$$Y(k) = \text{DFT}\{y_w[n]\} \tag{2.10}$$

- Linear filterbank: Next, a linearly spaced filterbank is employed. This means that the filters are equally spaced in the linear frequency domain. Let $H_i(k)$ represent the i^{th} filter in the linear filterbank.

- Compute filterbank energies: The power spectrum $|Y(k)|^2$ is then mulitplied with the fre-

quency response of each filter in the linear filterbank. The output of each filter is then summed to obtain the filterbank energies as illustrated in Equation 2.11 :

$$E_i = \sum_{k=1}^{N} |Y(k)|^2 \cdot H_i(k) \tag{2.11}$$

where E_i is the energy output of the i^{th} filter and N is the number of bins in the uniform filter.

- Logarithm: The logarithm of the filterbank energies is taken to obtain the log filterbank energies as:

$$L_i = \log(E_i) \tag{2.12}$$

- Discrete cosine transform: Finally, the discrete cosine transform is applied to the filterbank energies to obtain the cepstral coefficients as given in Equation 2.13.

$$c_j = \sum_{i=1}^{N} L_i \cdot \cos\left(j\left(i - \frac{1}{2}\right)\frac{\pi}{N}\right), \quad j = 1, 2, \ldots, M \tag{2.13}$$

where c_j is the j^{th} cepstral coefficient, N is the number of linear filters, M is the number of cepstral coefficients to be computed and L_i is the log-energy output of the i^{th} filter.

These c_j values are referred to as the LFCC features.

2.4.3 Gaussian Mixture Model Classifier

In this section, we describe the Gaussian mixture model (GMM) based back-end that is used as a classifier to develop the baseline replay attack detection (RAD) systems. A GMM (λ) is a parametric probability density function and is mathematically defined as the weighted sum of N Gaussian component densities as shown in the equation below [76]:

$$p(x|\lambda) = \sum_{i=1}^{N} w_i g(x|\mu_i, \Sigma_i) \tag{2.14}$$

where x is D-dimensional feature vector, w_i and $g(x|\mu_i, \Sigma_i)$ are the mixture weight and component density for the i^{th} Gaussian component respectively, μ_i is the mean vector and Σ_i is the covariance matrix.

The i^{th} component density is a D-variate Gaussian function which can be expressed as follows:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{\frac{1}{2}}} \exp\Big\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\Big\} \tag{2.15}$$

where w_i, μ_i and Σ_i are called the parameters of the GMM (λ). Hence, the GMM is denoted as:

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, 3, ..., N \tag{2.16}$$

Given a set of feature vectors $X = \{x_1, x_2, x_3, ..., x_L\}$ and the configuration of the GMM, the process of training the GMM involves estimating the parameters $\{w_i, \mu_i, \Sigma_i\}$. There are several techniques to estimate these parameters. Maximum likelihood (ML) estimation and maximum aposteriori (MAP) estimation are two such widely used techniques [76], [77]. In this thesis, we use ML estimation for GMM parameter estimation.

The GMM likelihood of a set of training feature vectors $X = \{x_1, x_2, x_3, ..., x_L\}$, assuming independence between feature vectors, can be expressed as:

$$p(X|\lambda) = \prod_{i=1}^{L} p(x_i|\lambda) \tag{2.17}$$

The goal of ML estimation is to find the values of the model parameters that maximize the likelihood of the GMM given the training data.

Taking the logarithm on both sides of Equation 2.17, the log-likelihood of the GMM model λ for the set of feature vectors $X = \{x_1, x_2, x_3, ..., x_L\}$ is given by:

$$\log p(X|\lambda) = \sum_{i=1}^{L} \log p(x_i|\lambda) \tag{2.18}$$

where $p(x_i|\lambda)$ is computed as shown in Equation 2.14.

2.5 Experimental Setup

In this thesis, two baseline RAD systems are developed for contrast purposes. The first baseline system is developed using CQCC features and a GMM back-end. The second baseline system uses LFCC features with a GMM back-end. The experimental details about these features and the back-end are provided in the subsequent sections. The experimental details of the ASV system necessary to calculate the t-DCF metric is also provided in this section.

2.5.1 CQCC Feature

The CQCC features for the baseline RAD system employ a constant-Q transform (CQT), with a maximum frequency capped at half the sampling frequency ($f_s = 16$ kHz). The minimum frequency is established at 15 Hz, nine octaves below the maximum frequency. Each octave consists of 96 bins. The resultant CQT spectrogram, scaled geometrically, undergoes resampling to a linear scale with a sampling period of 16. Subsequently, a discrete cosine transform (DCT) is applied to generate a set of static cepstral coefficients. This comprehensive set of CQCC features encompasses 29 static coefficients along with their respective delta and delta-delta coefficients, calculated from the comparison of two consecutive frames. The code for extracting the CQCC feature was made available as part of the ASVSpoof 2017 challenge and can be found at this link[1].

2.5.2 LFCC Feature

The LFCC baseline utilizes a short-term Fourier transform method. It begins by segmenting the input signal into frames, each with a duration of 20 milliseconds, and applying a Hamming window of the same length with a shift of 10 milliseconds. Subsequently, the power magnitude spectrum of each frame is determined using a 512-point fast Fourier transform (FFT). Following this, a triangular filterbank with 20 channels, evenly spaced, is employed to derive a set of 20 coefficients. LFCC features are then derived by computing the DCT of the log-energies output of the linear filterbank. These features consist of 19 static coefficients plus the zeroth coefficient, and the corresponding delta and delta-delta coefficients computed using two adjacent frames.

2.5.3 GMM Back-end

To develop the RAD systems, we first extract the features from the genuine and replayed speech signals of the training subset of the data. The genuine and replayed features are used to train two GMMs (a genuine model λ_{gen} and a replay model λ_{rep}) of 512 components each. Diagonal covariance matrices are used in the two models. The parameters of the GMMs are estimated using the expectation maximization (EM) algorithm. The EM algorithm is initialized with the help of k-means algorithm. Each iteration of the EM algorithm refines the parameters and we run it for

[1]https://www.asvspoof.org/index2017.html

100 iterations. This completes the process of training the two models. Next, the features of each test trial are extracted. This results in a set of feature vectors S for each trial. The log-likelihood score $\Lambda(S)$ for a test trial is then calculated in the following manner:

$$\Lambda(S) = \log p(S|\lambda_{gen}) - \log p(S|\lambda_{rep}) \tag{2.19}$$

The log-likelihood scores for the entire test set are then used to compute the performance metrics as described in Section 2.2.7 and Section 2.3.5.

2.5.4 ASV System

The task with the ASVSpoof 2019 database is to develop both a RAD system and an ASV system. The performance of the RAD system can be assessed with the EER metric. On the other hand, the combined assessment of the two systems which include genuine, impostor, and spoofed trials is done in terms of the t-DCF metric. Thus, to evaluate the performance of a RAD system on the ASVSpoof 2019 database, it is necessary to build an ASV system as well. However, since the focus of this thesis is only on the development of RAD techniques, the ASV system has not been built. Instead, the scores generated by the ASV system developed by the organizers of the ASVSpoof 2019 challenge have been used to benchmark the RAD systems. The details of this ASV system are provided below.

The ASV system utilizes DNN-based x-vector speaker embeddings [78] in conjunction with a probabilistic linear discriminant analysis (PLDA) back-end [79]. The x-vector extractor is a pre-trained neural network [2] accessible within the Kaldi toolkit [80]. It is trained using MFCC features extracted from audio data sourced from 7325 speakers from the VoxCeleb1 and VoxCeleb2 databases [81]. The x-vector model comprises a 5-layer deep time-delay neural network (TDNN), followed by statistics pooling and two fully connected layers before a softmax output. The statistics pooling layer transforms the TDNN output from frame-level to utterance-level representations by computing the mean and standard deviation of features over time. The x-vector embeddings are derived from the first fully connected layer after the pooling layer and those are 512 dimensional. These embeddings are extracted without the application of the rectified linear unit (ReLU) activa-

[2] http://kaldi-asr.org/models/m7

tion function or batch normalization. The network is trained using a stochastic gradient descent algorithm. More detailed information on network parameters and data preparation are found in [78].

For each enrolled speaker, the x-vector representations of their enrollment utterances were combined by averaging them, resulting in a single x-vector per speaker. Prior to the scoring process based on the log-likelihood ratio using PLDA, the x-vectors underwent several preprocessing steps. Firstly, they were centered and then they were reduced to 200 dimensions using a linear discriminant analysis transform. This transformation helps to whiten the within-class covariance matrix. Finally, the x-vectors were normalized to have a unit length. The Kaldi implementation of PLDA is then utilized for the scoring process [3].

2.6 Results and Discussion

In this section, the performances of the baseline RAD systems on the two databases are reported. The results of the experiments conducted using the ASVSpoof 2017 v2.0 database are given in Table 2.3 in terms of minimum DCF and EER. This table shows the baseline performances on both the development and evaluation sets of the database. The LFCC-GMM-based RAD system results in an EER of 17.11% while the CQCC-GMM-based RAD system gives an EER of 9.19% on the development set. On the evaluation set, the CQCC and LFCC-based systems provide EERs of 16.89% and 13.84%, respectively. Thus, it can be inferred that the CQCC features perform better than the LFCC features for RAD.

A similar trend can be seen in the results of the experiments done on the PA set of the ASVSpoof 2019 database. The performances of the baseline RAD systems on this database are provided in Table 2.4. The LFCC and CQCC-based baseline systems yield EERs of 13.54% and 11.66% on the evaluation set of the database. The joint performance of a RAD and an ASV system is also given in the table in terms of t-DCF. The CQCC-based system combined with the ASV system gives a t-DCF of 0.261 whereas the LFCC-based system coupled with the ASV system results in a t-DCF of 0.301. It is clear that the CQCC features are performing better than the LFCC features on both databases.

[3]https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

Table 2.3: Performances of baseline systems on the ASVSpoof 2017 v2.0 database

Feature	Development		Evaluation	
	min-DCF	EER (%)	min-DCF	EER(%)
LFCC	0.703	17.11	0.813	16.89
CQCC	0.455	9.19	0.739	13.84

Table 2.4: Performances of baseline systems on the PA set of the ASVSpoof 2019 database

Feature	Development		Evaluation	
	t-DCF	EER (%)	t-DCF	EER(%)
LFCC	0.255	11.96	0.301	13.54
CQCC	0.207	10.39	0.261	11.66

2.7 Conclusion

In this chapter, we delved into the fundamental concepts and methodologies employed in the development of baseline RAD systems, focusing on feature extraction techniques and back-end classifiers. We explored two baseline features: CQCC and LFCC, each offering unique advantages for speech signal analysis and classification. The CQCC features leverage the CQT, which adapts frequency resolution according to human auditory perception. This adaptability provides high-frequency resolution at lower frequencies and high temporal resolution at higher frequencies, making CQCC particularly effective for audio signal processing. For the classification task, a GMM back-end was employed. The parameters of the GMM were estimated using the EM algorithm, with maximum likelihood estimation optimizing the model fit to the training data. The performances of these systems were evaluated using the ASVSpoof 2017 v2.0 and ASVSpoof 2019 databases. The CQCC-GMM-based RAD system consistently outperformed the LFCC-GMM system, achieving lower EERs on the development and evaluation sets of both databases. Additionally, the combined performance of a RAD system with an ASV system, assessed using the t-DCF metric, further demonstrated the benefit of CQCC features over LFCC features.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

3

Handcrafted Pitch-Synchronous Source Features

Contents

3.1 Introduction

In the existing literature, the majority of RAD features attempt to capture the replay attack information by analyzing the signal spectrum or its smoothed version. As discussed in Chapter 1, the speech signals can be decomposed into source and filter components. We hypothesized that both these components would carry information about replay attacks. So far, the RAD approaches have targeted the filter component of speech only. Little research has been done on the role of the source component in detecting replay attacks. This chapter begins with a discussion of the different techniques found in the literature to model the speech source. Following this, the preliminary studies undertaken to understand how replay attacks impact the speech source component are presented. In these studies, we try to discover the differences in the nature of the source signal of genuine and replayed speech. Two different representations of the speech source signal are considered for these studies. They are called the zero frequency filtered (ZFF) signal and the linear prediction residual (LPR) signal. First, a visual comparison is made between the ZFF and LPR signals of pairs of genuine and replayed speech signals to appreciate the differences in the source representations. Then two handcrafted features are defined that can capture the dissimilarities in the source representations. One of these features is a two-dimensional feature containing the epoch intervals and the corresponding strength of excitation extracted from the ZFF signal. The mean and the skewness of the peak-to-sidelobe ratio (PSR) of the Hilbert Envelope of the LPR are taken as the second source feature. These features are extracted for each pitch period and hence are known as pitch-synchronous features.

Apart from these two source features, this chapter also explores two more cepstral features for RAD, in addition to the baseline LFCC and CQCC features. These features are the instantaneous frequency cosine coefficient (IFCC) feature [82] and the mel frequency cepstral coefficient (MFCC). Individual RAD systems are developed using the source and cepstral features as the front-end and GMM as the back-end on the ASVSpoof 2017 v2.0 database. The performances of the RAD systems are reported for both the development and evaluation sets of this database.

The rest of this chapter is organized as follows. The review of glottal source modeling techniques is presented in Section 3.2. The process of extracting epochs from the ZFF signal is given in Section 3.3.1.1. In Section 3.4, the features used for RAD are explained in detail. Section 3.5

describes the experiments conducted for this chapter and describes the process of development of the RAD systems. The results of these experiments and the following discussion are reported in Section 3.6. Conclusions drawn from this chapter are presented in Section 5.6.

3.2 Review of Glottal Source Modelling

Glottal source or voice source refers to the volume velocity waveform that represents the excitation caused by the vibration of the vocal folds [83]. This thesis explores the impact of replay attacks on the glottal source signal. An attempt is made to study the significance of the glottal source signal and its different parameters for replay attack detection. Hence, accurate estimation of the glottal source from the speech signal is an indispensable aspect of the proposed methodology. In this preliminary section, we present a brief review of the various glottal flow estimation techniques that have been reported in the literature.

The linear model of speech production posits that speech is generated when the voice source signal is filtered through the vocal tract transfer function [84]. Several models for glottal source have been presented which includes the classical Liljencrants–Fant (LF) model [85]. The Rosenberg [86], Fujisaki-Ljungqvist (FL) [87] and Rosenberg++ [88] models are some of the other famous glottal source models.

Most glottal flow or glottal source estimation is performed with the help of an inverse filtering process. These techniques involve creating a parametric model of the vocal tract filter first and then passing the speech signal through an inverse filter corresponding to the estimated vocal tract transfer function. The resulting signal is referred to as the glottal source signal. Inverse filtering techniques can be divided into two categories based on the method of estimating the vocal tract transfer function. In the first category, this estimation is performed in the closed phase of glottis while in the second category, it is done using an iterative/adaptive process. Linear prediction analysis is performed to estimate the transfer function using an all-pole filter in most of the methods. One such closed phase inverse filtering method that is widely used is given in [89]. A popular iterative method is the iterative adaptive inverse filtering proposed in [90].

Joint source-filter optimization has also been utilized to perform inverse filtering where the LF and Rosenberg models are used to represent the glottal source signal [91]. Glottal inverse

filtering (GIF) methods have also been developed using a combination of causal (minimum phase) and anticausal (maximum phase) components of the speech signal. The zeros of the z-transform method [92], [93] and the complex cepstrum decomposition method [94] happen to be two different approaches in this category.

3.3 Representations of Glottal Source Signal

In this chapter, we focus on the development of handcrafted source features for RAD. To investigate the differences in the genuine and replayed voice source signals, two different representations are used to approximate the voice source signal. The first representation is the ZFF signal and the second is the LPR signal. A detailed description of these two representations is presented in this section.

3.3.1 Zero Frequency Filtered Signal

Speech is produced by stimulating the changing vocal tract system through three types of excitation: glottal vibration, frication, and burst. The main type of excitation is glottal vibration. Though excitation occurs throughout speech production, it is most significant during glottal vibration, especially when there is high energy in a brief period, resembling an impulse. These impulse-like traits typically appear at the moments of glottal closure in each glottal cycle. These traits suggest that vocal tract excitation can be modeled as a sequence of impulses with varying intensities. The impulse-like excitation causes a uniform frequency range discontinuity, including at zero frequency. This means that even a zero-frequency filter should capture information about these discontinuities. The benefit of using a zero-frequency filter is that its output is unaffected by the vocal-tract system's resonances, which occur at much higher frequencies.

An ideal zero-frequency resonator is utilized to filter the speech signal. This resonator is a second-order infinite impulse response (IIR) filter with real poles on the unit circle. Two ideal zero-frequency resonators in series characterize the discontinuities from impulse-like excitation in voiced speech. This cascade provides a 24 dB per octave roll-off, effectively eliminating high-frequency components beyond zero-frequency. Filtering a speech signal twice through a zero-frequency resonator produces an output that grows or decays polynomially over time. The large direct current (DC) offset or

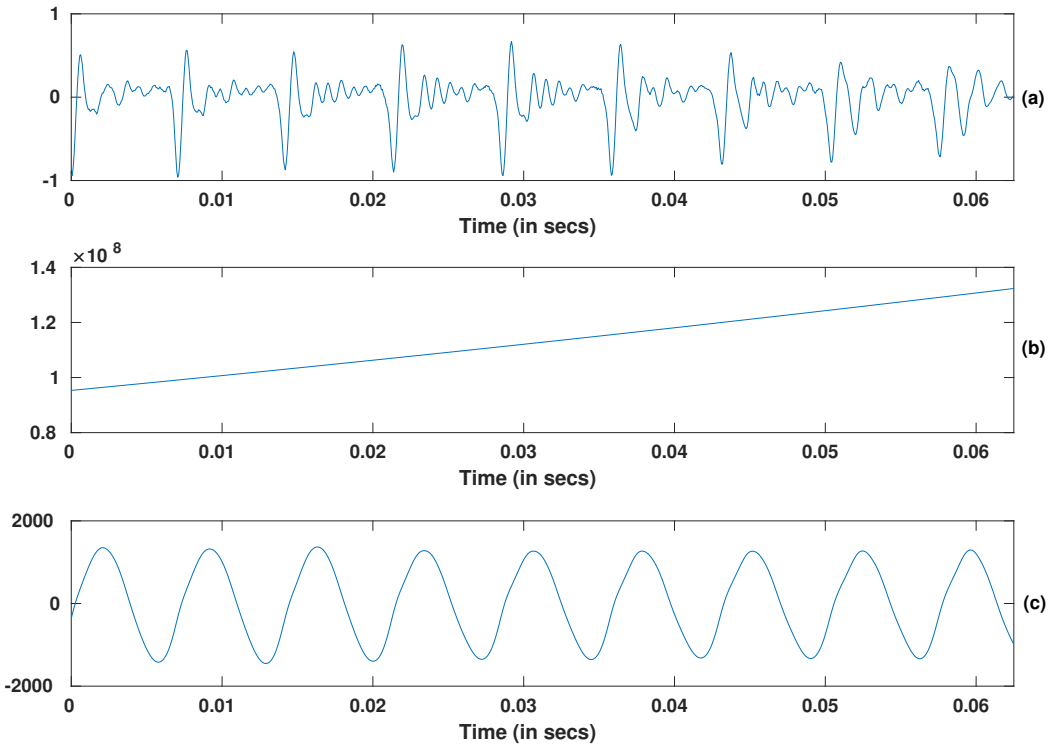

Figure 3.1: Figure showing the steps in calculating the ZFF signal. (a) A voiced segment of a speech signal, (b) the output obtained by passing the speech segment in (a) through a 0-Hz resonator twice, and (c) the mean-subtracted output of cascaded 0-Hz resonators or the ZFF signal.

bias resulting from zero-frequency filtering can overshadow the effects of impulse-like excitation. Subtracting the local mean, calculated over a small window, can highlight the discontinuities. A window size of one to two times the average pitch period is sufficient for local mean subtraction. This mean-subtracted signal is known as the zero-frequency filtered (ZFF) signal. The stages involved in extracting the ZFF signal from a speech signal are depicted in Figure 3.1.

3.3.1.1 Method of Epoch Extraction

Analyzing voiced speech involves understanding the frequency response of the vocal tract system and the glottal pulses that act as the excitation source. Although these glottal pulses drive the excitation for voiced speech, the critical excitation happens within each glottal pulse, particularly at the moment of glottal closure, known as the epoch. Accurate estimation of epoch locations is essential for many speech analysis tasks. After glottal closure, the glottal airflow often drops to zero, causing the supralaryngeal vocal tract to acoustically separate from the trachea. Consequently, the

speech signal during the closed phase reflects the natural resonances of the supralaryngeal vocal tract system. Analyzing the speech signal in these closed-phase regions allows for precise estimation of the frequency response of supralaryngeal vocal tract system [95], [96]. By identifying the epochs, one can determine the voice source characteristics through a detailed analysis of the signal within each glottal pulse. The regions of the speech signal immediately following epochs are more resistant to external degradation due to the significant excitation. These moments of significant excitation also play a crucial role in human perception.

The ZFF signal shows rapid changes around the positive zero crossings, which can be used to identify epochs. Interestingly, for impulse sequences, even aperiodic ones, the positive zero-crossing instants correspond to the impulse locations. However, for random noise excitation in a time-varying all-pole system, there is no similar relationship between the excitation and the filtered signal. Additionally, the filtered signal has much lower values for random noise excitation compared to impulse sequence excitation. Figure 3.2 illustrates the epochs extracted from the ZFF signal. It first depicts a segment of a speech signal. This is followed by the DEGG signal with the ground truth markings of the epoch locations. Finally, the positive zero crossings of the ZFF signal are identified as epochs. It can be observed that the identified epochs match closely with the actual epochs.

3.3.2 Linear Prediction Residual Signal

Linear predictive coding (LPC) is a widely used speech processing method and it closely approximates the source-filter model of speech production. The glottis (middle part of the larynx which locates vocal folds) forms the source. It produces a buzz that is characterized by its frequency (pitch) and intensity (loudness). The tube-like structure formed by the throat and mouth is referred to as the vocal tract and is characterized by its resonances; these resonances give rise to formants. The LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the residue. The process of removing the formants is called inverse filtering, and the residue after the subtraction of the modeled filter response characterizes the excitation or source signal. This residue is referred to as the linear prediction residual (LPR) signal.

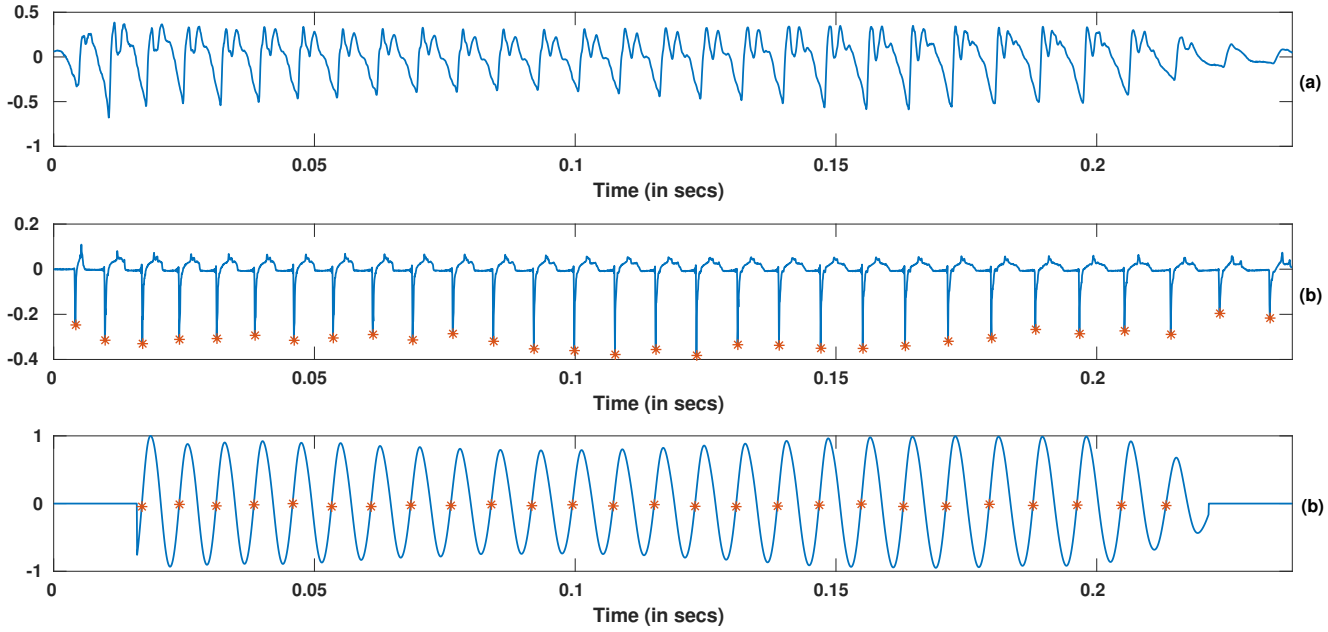

Figure 3.2: Figure showing the detected epoch locations. (a) A voiced segment of a speech signal, (b) differenced electro-glottal graph (DEGG) signal marked with the ground truth of epoch locations, and (c) ZFF signal and the detected epoch locations marked with red stars.

The LP analysis works on the principle that a sample value in a correlated, stationary sequence can be predicted as a linear weighted sum of the past p samples. If $s(n)$ denotes a sequence of speech samples, then the predicted value at the time instant n is given by the equation

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \; s(n-k) \tag{3.1}$$

where $a_k,\ k = 1, 2, \ldots, p$ is the set of linear predictor coefficients and p is the order of the LP filter. The error at time n and the sum of squared errors E are given by

$$r(n) = s(n) - \ \hat{s}(n) \tag{3.2}$$

The error signal $r(n)$ obtained by inverse filtering the speech signal is the LPR signal. The LPR signal has large error values at regular intervals and can be attributed to the periodic impulses of excitation. Hence the LPR is a good approximation to the excitation source signal and can be used further to extract the excitation source characteristics

3.4 Handcrafted Features for Replay Attack Detection

The ZFF and LPR signals are representations of the voice source. Initially, a visual analysis of these representations for the genuine and replayed speech signal is performed. This analysis revealed that the source signal representations for the replayed speech are distorted compared to genuine speech due to the presence of replay channel artifacts. To encode this information, two handcrafted source features are designed. The first one is derived from the ZFF signal and is called the epoch feature (EF). It is a two-dimensional feature that captures the differences in the epoch intervals and epoch strength between the genuine and replayed source signal. The second feature is computed from the LPR signal and is called the peak-to-side lobe ratio mean and skewness (PSRMS). In this section, a detailed explanation of the handcrafted features is provided.

3.4.1 Epoch Feature

The main focus of this work is to understand the dissimilarities in a genuine and a spoofed speech signal. To this end, the first source feature used is epoch instants and their corresponding epoch strength. Epochs are defined as the instants where significant excitation is present during speech production [97]. Glottal closure instants are the regions around which the most significant excitations occur for voiced speech. In this work, the epochs are extracted using the zero frequency filter (ZFF) method [97]. The following steps are used to determine epochs and epoch strengths from the speech signal $s[n]$ [98].

- Difference the speech signal

$$x[n] = s[n] - s[n-1] \tag{3.3}$$

- Pass $x[n]$ twice through the zero frequency resonator.

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{3.4}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \tag{3.5}$$

where, $a_1 = -2$ and $a_2 = 1$. This is equivalent to integrating four times successively.

- Remove the trend by subtracting $y_2[n]$ with the average value of $y_2[n]$ calculated over the window length of the average pitch period.

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \tag{3.6}$$

where, $2N+1$ is the number of samples in the average pitch period. This trend-removed signal is called the ZFF signal.

- The positive crossings of the ZFF signal are taken as the epochs.

- Slope of the ZFF signal is called the epoch strength or strength of excitation $S_e(k)$ [99].

$$S_e(k) = \big|y[k+1] - y[k]\big| \tag{3.7}$$

where k represents the k^{th} epoch location.

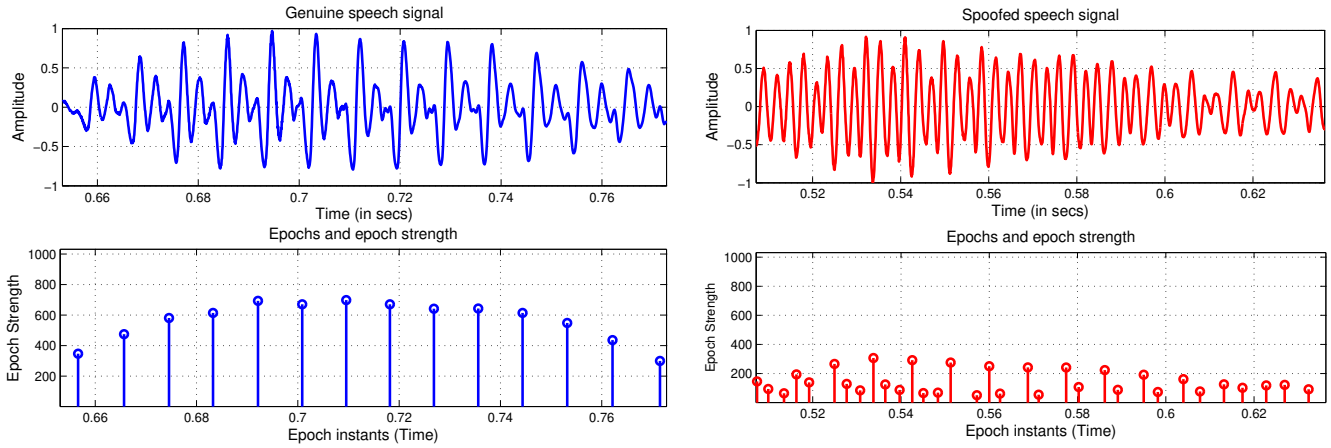

Figure 3.3: Figure showing the epochs and their strength for a segment of a genuine and the corresponding replayed speech signal

Figure 3.3 shows the plot of epoch and epoch strength for a speech segment. It is observed that the epochs are equally spaced for the genuine speech while there is no such structure in the spoofed speech. The epoch strengths are also higher for genuine speech as compared to spoofed speech. Thus, a two-dimensional feature containing the difference of epochs (epoch interval) as one dimension and the corresponding epoch strength as another dimension is created which is referred to as the epoch feature (EF). In Figure 3.4, the distribution of EF for a genuine speech and its corresponding spoofed speech are shown which highlights their discrimination.

Then the glottal activity regions (GAR) in the speech are calculated and the corresponding strengths of excitation occurring only within GARs are considered. The method used for extracting the GARs for this thesis is explained in detail in Appendix A.

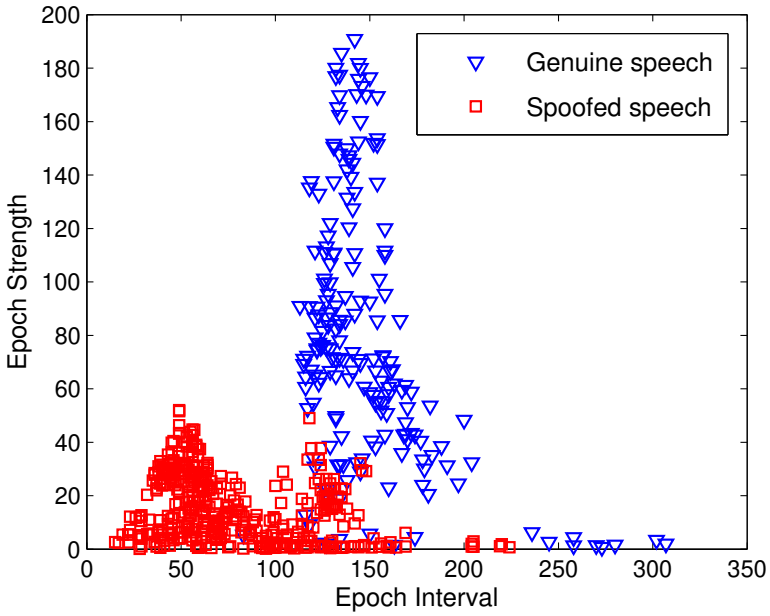

Figure 3.4: Figure showing discrimination achieved with EF feature for a genuine speech signal and its replayed version

3.4.2 Peak-to-Side Lobe Ratio of the Hilbert Envelope of the LP Residual

The LP residual gives information about the excitation source information, most importantly the epoch sequence for a segment of voiced speech. The residual error is large around the epochs and the prediction is poor [100]. However, since the residual signal amplitudes depend on the phase of the signal it may cause ambiguity in determining the epochs. Thus, instead of using the LP residual directly, the Hilbert envelope of the LP residual signal is used which helps in reducing the ambiguity about the peaks [100]. The Hilbert envelope $h(n)$ of the LP residual $r(n)$ is computed using the following equation.

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{3.8}$$

where, $r_h(n)$ is the Hilbert transform of $r(n)$. Figure 3.5 shows the Hilbert envelope of the LP residual for a segment of a genuine and a spoofed speech. It is observed that the peaks in the

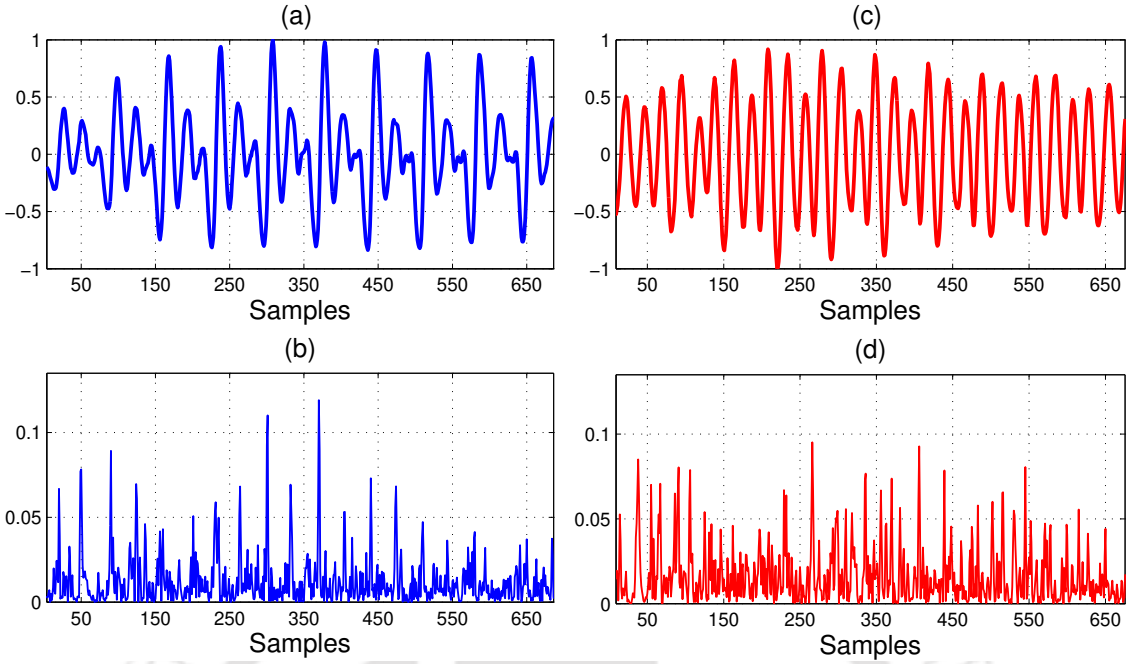

Figure 3.5: (a) A segment of a genuine speech example in ASVSpoof 2017 v2.0 database, (b) Hilbert envelope of the LP residual of that genuine speech segment, (c) the matching segment of replayed speech corresponding to the chosen genuine speech, and (d) Hilbert envelope of the LP residual of that replayed speech segment.

genuine speech are more well-defined than the peaks in the spoofed speech and are less affected by the side lobes. Hence, we use the parameter PSR of the Hilbert envelope of the LP residual.

The computation of the PSR parameter requires the knowledge of the epoch locations. These epoch locations are calculated using the ZFF signal as explained in Section 3.3.1.1. Once the epoch locations are obtained, the peaks are searched in the Hilbert envelope of the LP residual within a window of 3 ms around the epoch locations. The typical pitch of adult human speech is between 100 Hz and 300 Hz. A 3-ms window (corresponding to a pitch of around 300 Hz) is chosen because it is expected that there will be no other epochs within this period. The window is small enough to avoid interference with adjacent epoch locations, yet wide enough to accommodate the expected variation in epoch location due to slight inaccuracies in ZFF signal computation.

The maximum peak value within this 3-ms window is taken as the peak of the Hilbert envelope of the LP residual of the speech signal. For calculating side-lobe values, the mean of sample values 1.5 ms to the right and 1.5 ms to the left of the peak value is taken. PSR is calculated by dividing the peak of Hilbert envelope of LP residual by the side-lobe value [101]. The histogram of the PSR

mean for the train set of the ASVSpoof 2017 database is shown in Figure 3.6. From this figure, it can be observed that the PSR mean of the genuine speech is much higher than that of the spoofed speech. The distribution is also more skewed for the spoofed speech as compared to that of the genuine speech. Taking these factors into consideration, a two-dimensional feature vector consisting of the mean and skewness of the PSR values of a signal is created and is referred to as PSRMS.

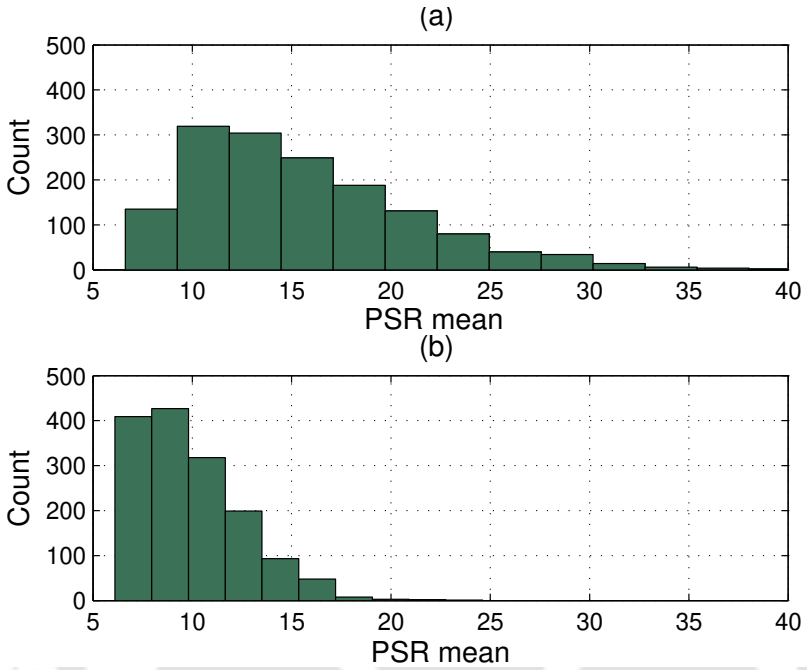

Figure 3.6: Histogram of the PSR mean values of (a) genuine utterances and (b) replayed utterances in the training set of the ASVSpoof 2017 v2.0 database

3.4.3 Instantaneous Frequency Cosine Coefficients

The instantaneous frequency cosine coefficients (IFCC) is an attempt to extract features from the analytic phase of speech signal for the speaker verification [82]. In order to overcome the problem of phase warping, the instantaneous frequency (IF) is computed with the help of Fourier transform properties without explicit involvement of computation of the analytic phase. The narrow-band components of speech are taken to compute IF in the following way,

$$\theta'[n] = \frac{2\pi}{N}\Re\left(\frac{F_d^{-1}kZ[k]}{F_d^{-1}Z[k]}\right) \tag{3.9}$$

where, F_d^{-1} denotes inverse discrete Fourier transform (IDFT), N being the length of the narrow band signal and $Z[k]$ is the DFT of the analytic signal $z[n]$, obtained from the narrow-band component of speech signal as explained in [102].

The computation of IF is followed by discrete cosine transform (DCT) on deviations in IF computed from narrow-band components of speech to extract IFCC features [82].

3.4.4 Mel Frequency Cepstral Coefficients

MFCCs are extensively utilized as spectral features in speech processing because they effectively capture the timbral aspects of a signal [103]. The calculation of MFCCs involves computing the spectrum of the speech signal using a DFT, transforming the Fourier coefficients into the mel scale, applying a logarithmic function, and using a DCT to remove redundant information [75]. The key components of MFCCs are the initial DCT coefficients that describe the general spectral shape. The first coefficient represents the average power of the spectrum, while the second coefficient approximates the broad spectral contour, corresponding to the spectral centroid. Higher-order coefficients provide more detailed spectral information [103]. The primary advantage of MFCCs lies in their efficient encoding of phonetic features and acoustic features such as speaker and environmental conditions. The MFCCs are calculated as follows:

$$\mathrm{MFCC}_i = \sum_{k=1}^{N} X_k \cos\left[i \left(k - \frac{1}{2} \right) \frac{\pi}{N} \right], \quad i = 1, 2, \ldots, M \tag{3.10}$$

where N is the number of triangular bandpass filters, M is the number of cepstral coefficients, X_k, $k = 1, 2, \ldots, N$ is the log-energy output of the k^{th} filter.

3.5 Experimental Setup

This section explains the experimental setups of the different RAD systems developed using the features explained earlier, and a summary of these systems is provided in Table 3.1.

3.5.1 EF-based RAD System

For each speech file, the glottal activity regions (GAR) are detected. EF feature is calculated within the GARs. Delta and double delta coefficients are computed from the EF features to obtain a 6-dimensional feature vector for each epoch. Z-score normalization is performed over these features. These features are then used to train two Gaussian mixture models of 128 components for genuine speech and spoofed speech.

3.5.2 PSRMS-based RAD System

In this system, the mean and skewness of PSR values are used as features. The given speech signal is first down-sampled to 8 kHz and then 12^{th} order linear prediction (LP) analysis is performed with a frame size of 20 ms and frameshift of 10 ms to obtain the LP residual signal. The Hilbert envelope of the LP residual signal is then calculated. The PSR values are determined from this envelope. Only the PSR values within the GARs are considered. The mean and skewness of the PSR values are then computed. These feature vectors are used to train two GMMs of 16 components each.

3.5.3 IFCC-based RAD System

To build this system, a short-time analysis of speech is done using Hamming window of duration 20 ms with a frameshift of 10 ms. Energy-based voice activity detection is performed to separate the speech regions from the silence regions. The narrow band components of speech are then taken to compute instantaneous frequency (IF) and then discrete cosine transform is applied to it to extract instantaneous frequency cosine coefficient (IFCC) features of 20 dimensions as a compact representation. The delta and delta-delta features are calculated to obtain 60-dimensional features. Two GMMs of 512 components each are trained on the genuine and spoof speech files. The log-likelihood ratio scores are used as test scores.

3.5.4 MFCC-based RAD System

The speech files are short-time processed with a frame size of 20 ms and frameshift of 10 ms. Energy-based voice activity detection is done to find the speech and non-speech regions. The speech regions are used to calculate 13-dimensional MFCC features excluding the energy coefficient. Delta and delta-delta coefficients are then calculated and appended to obtain a 39-dimensional feature vector for each speech frame. Two GMMs of 512 components each are built on the MFCC feature vectors for genuine and spoof speech and likelihood ratios are used as test scores.

Table 3.1: The details of different feature-based RAD systems developed on the ASVSpoof 2017 v2.0 database.

RAD System	Feature dimensionality (composition)	Classifier
EF-based	6 (2-static + 2-Δ + 2-$\Delta\Delta$)	GMM (128 components)
PSRMS-based	2 (2-static)	GMM (16 components)
IFCC-based	60 (20-static + 20-Δ + 20-$\Delta\Delta$)	GMM (512 components)
MFCC-based	39 (13-static + 13-Δ + 13-$\Delta\Delta$)	GMM (512 components)

3.6 Results and Discussion

The performances of the different feature-based RAD systems developed and evaluated on the ASVSpoof 2017 v2.0 database are reported in Table 3.2 in terms of min-DCF and EER. The table also includes the performances of the baseline CQCC- and LFCC-based RAD systems. From the table, it can be observed that the EF-based RAD system results in an EER of 33.07% and 32.51% for the development and evaluation sets, respectively. For the PSRMS-based RAD system, the corresponding EERs are 31.6% and 28.9%, respectively. Thus, the PSRMS-based system provides better RAD performance as compared to the EF-based system. The IFCC-based RAD system does not generalize well to the evaluation set. The MFCC-based RAD system yields much better performance compared to the proposed handcrafted source features. A similar trend can be seen for the baseline CQCC- and LFCC-based RAD systems. Both the baseline systems outperform the two proposed source-based features by a significant margin. This can be attributed to the fact that these two handcrafted features are defined very naively as compared to the CQCC and LFCC features and hence fail to capture the intrinsic RAD information embedded in the speech signal. Furthermore, these source features are defined only at the instants of significant excitation, the GCIs, and hence capture very local information rather than utilizing the entire source signal. The baseline features on the other hand make use of the entire speech signal. As a result of that, the proposed handcrafted source features (EF and PRSMS) are unable to compete with the CQCC and LFCC features. However, given that two naive handcrafted features extracted only from regions around the pitch periods result in decent RAD performance, it does convey the fact that considerable replay attack information is present in the source signal.

Table 3.2: Performances of different feature-based stand-alone RAD systems developed and evaluated on the ASVSpoof 2017 v2.0 database.

RAD System	Development		Evaluation	
	min-DCF	EER (%)	min-DCF	EER (%)
EF-based	0.997	33.07	1.000	32.51
PSRMS-based	0.974	31.60	0.992	28.90
IFCC-based	0.841	24.81	1	35.19
MFCC-based	0.724	18.90	0.864	23.55
LFCC-based	0.703	17.11	0.813	16.89
CQCC-based	0.455	9.19	0.739	13.84

3.7 Conclusion

This chapter lays the foundation for the work done in this thesis. The main objective is to study the nature of genuine and replayed speech source signals and to exploit their characteristics so that some source features can be defined for RAD. It begins with an introduction to speech source modeling techniques. Then two handcrafted source features are proposed for RAD. These two features are named EF and PSRMS. Apart from these source features, this chapter also deals with two more features namely IFCC and MFCC for RAD. Using these features and GMM back-ends different RAD systems are developed on the ASVSpoof 2017 v2.0 database. It is noted that the EF and PSRMS based RAD systems provide decent performance. However, the baseline CQCC and LFCC based RAD systems yield significantly better results both in terms of EER and min-DCF on the above mentioned database. This is due to the fact that the CQCC and LFCC features are extracted from the entire source signal and hence can take advantage of replay information present in the whole signal. On the other hand, the EF and PSRMS features use information present only around the GCIs in the source signal and hence most of the replay information present in the signal is discarded. Nevertheless, the source features result in reasonable RAD performance and hence it confirms our hypothesis that source signals also contain substantial replay information. Thus, the source-level features can serve as a viable alternative representation for replay attacks.

As discussed earlier, the proposed source features suffer from the drawback that they extract information only around the GCIs of the source signal because of the way in which they are defined. This is a major impediment in achieving more competitive performance as a large part of the source signal is discarded in the processing of these features. Also, the ZFF signal filters out most of the

information as its purpose is to find out only the GCIs. The LP residual does contain information in the entire signal but there are multiple bipolar peaks around the epoch which makes the task of unambiguous GCI detection difficult. To overcome these drawbacks, a new representation of the source signal and a new processing of that signal is proposed in Chapter 4.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

4

Transform-based Pitch Synchronous Source Features

Contents

4.1 Introduction

In the previous chapter (Chapter 3), two source features namely EF and PSRMS were explored that characterize the excitation source behavior around the glottal closure instants (GCIs). Due to the nature of the processing involved in extracting these features, the information present in the rest of the source signal is not utilized. Moreover, the source signal representations used in the previous chapter are not found to be very effective for RAD. The ZFF signal contains information only about the GCIs and all other information is filtered out. The LPR signal, on the other hand, contains multiple bipolar peaks around the GCIs. To deal with these issues, a new representation of the source signal is utilized in this chapter. This representation is called the integrated linear prediction residual (ILPR) signal which models the temporal shape of the speech source signal between two GCIs. In the case of spoofed speech, the ILPR signal is also expected to contain the replay attack artifacts. Thus, the nature of the source dynamics between two GCIs for the genuine and replayed signals will be different. It is hypothesized that characterizing the source temporal dynamics between two GCIs will result in enhanced RAD performance compared to the handcrafted features proposed earlier.

To characterize the source dynamics between two GCIs, the segment of the ILPR signal between two GCIs is considered. However, the ILPR segment between two GCIs does not yield fixed dimensional vectors as the number of samples between any two GCIs happens to vary. In order to solve this problem, the ILPR signal is applied with discrete cosine transform (DCT)in a pitch synchronous manner. On account of the energy compaction achieved with DCT, a fixed dimension representation can be obtained. So derived compressed excitation source features are called compressed ILPR (CILPR) features. Since these features are extracted for every GCI and DCT is applied to them, they are termed transform-based pitch-synchronous features.

Using these features a RAD system is developed with a GMM back-end on the ASVSpoof 2017 v2.0 database. First, the system is evaluated on the development set of the database. Next, two different sets of experiments are performed for the evaluation set. The first set of experiments is conducted using only the training set of the database to learn the GMMs. In the second set, data from both training and development sets are taken to build the GMMs.

The remainder of the chapter is organized in the following way. An introduction to the ILPR

signal is provided in Section 4.2. Section 4.3 explains the method of extraction of the CILPR feature in detail. In Section 4.4, the process of development of the proposed RAD system using CILPR is described. Experimental results and discussions are provided in Section 4.5. An ablation study is presented in Section 4.6. Finally, the conclusions are given in Section 5.6.

4.2 Integrated Linear Prediction Residual

The voice source signal contains more information in the neighbourhood of the epochs. Hence, correctly identifying the epochs is an important step in the extraction of any source feature. The LPR signal was used to approximate the voice source signal in Chapter 3. LP-based inverse filtering was used to obtain the LPR signal. First, a pre-emphasized speech signal was used to estimate the LP coefficients and then the pre-emphasized signal was passed through the inverse filter. The output from the inverse filter is referred to as the LPR signal. However, the LPR signal contains multiple bipolar peaks around the epoch which makes the process of identifying epochs difficult. The existence of these multiple peaks is due to the application of pre-emphasis to the speech signal, a differencing operation, which boosts the high-frequency components. This problem of ambiguous peaks can be avoided if the inverse filtering is done directly on the input speech signal without pre-emphasis. The LP coefficients for the inverse filtering are however calculated from pre-emphasized Hanning windowed speech samples in this case. The resulting residual signal is further smoothed by the application of a 5-point symmetric moving average. This smoothed signal is referred to as the integrated linear prediction residual (ILPR) signal [104]. The smoothing operation further reduces the ambiguity in the peaks near the epochs. In this chapter, we propose the use of the ILPR signal as the representation of the voice source. Figure 4.1 shows the difference in the nature of the LPR and ILPR signals for a voiced segment of a speech signal. It can be observed that the LPR contains multiple peaks around the epoch instants. This is smoothened in the ILPR signal and the epochs are not corrupted by the presence of multiple peaks.

4.3 Capturing the Glottal Source Dynamics

Figure 4.2 shows four glottal cycles of a speech signal and the corresponding ILPR for the original and spoofed signal. It can be seen from the figure that the dynamics of the ILPR signal between

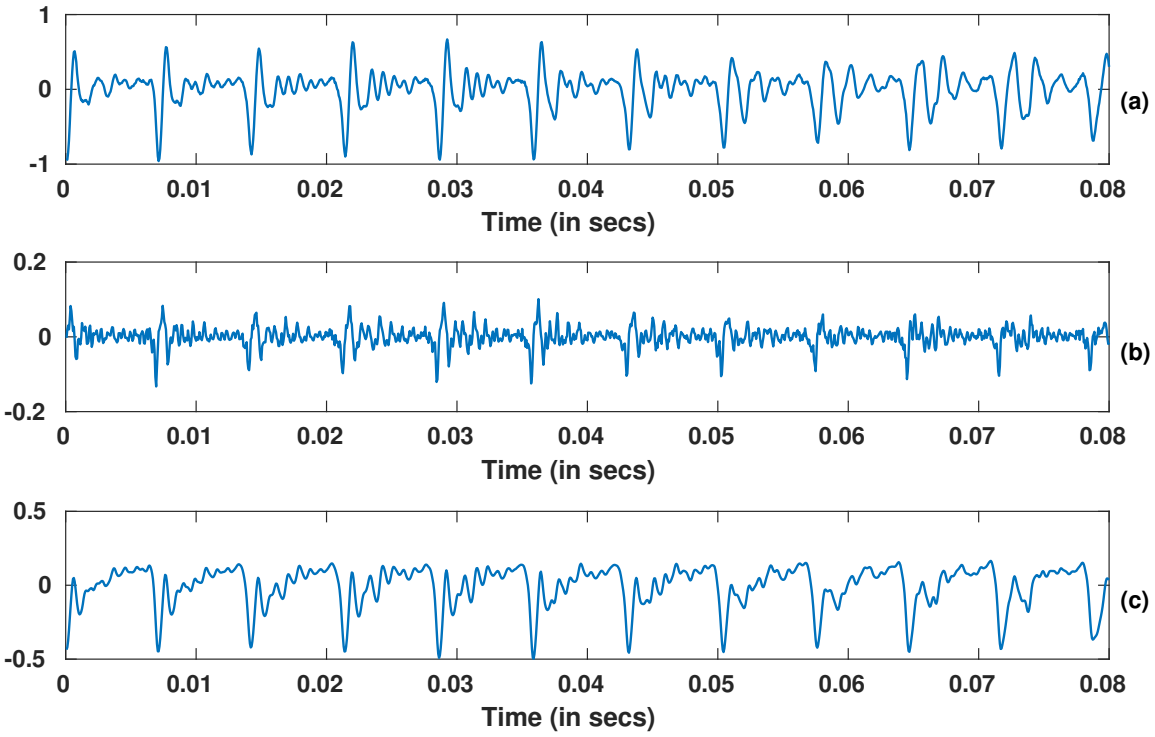

Figure 4.1: Figure depicting the difference between an LPR and ILPR signal. (a) voiced segment of a speech signal, (b) corresponding LPR of the voiced speech segment, (c) corresponding ILPR of the voiced speech segment

two GCIs are distorted from that of the original. The CILPR is computed from the ILPR-based voice source representation and captures the temporal shape of the voice source signal between two GCIs. This feature has also been explored for speaker identification in [105]. ILPR is estimated by passing a non-pre-emphasized version of the speech signal through an LP inverse filter, the LP coefficients of the inverse filters are obtained from the corresponding pre-emphasized speech signal. The LP order is considered to be $f_s/1000 + 4$, where, f_s is the sampling frequency. CILPR feature is computed pitch synchronously and requires GCIs to mark the pitch period. Let, $r_i(n)$ be a pitch synchronous segment of ILPR signal between i^{th} and $(i+1)^{th}$ GCIs. Then, DCT-II of $r_i(n)$ segment is computed by projecting it into the discrete cosine basis, as given below,

$$c(k) = \sum_{n=0}^{N-1} r_i(n) \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right], \tag{4.1}$$

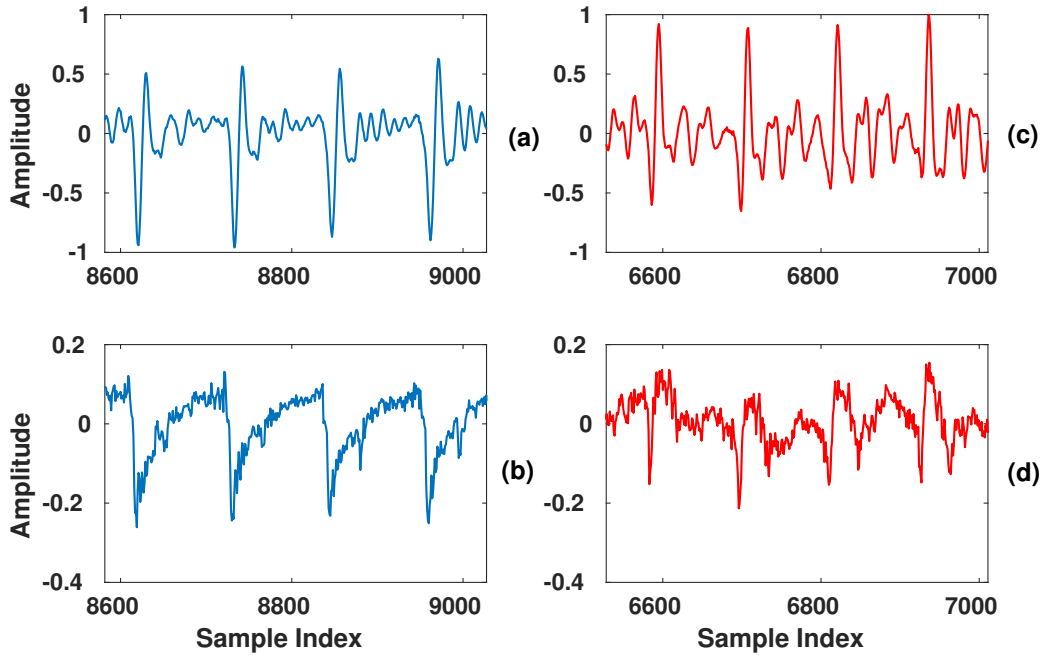

Figure 4.2: ILPR signals for segments of genuine and corresponding replayed speech signals. (a)-(b) and (c)-(d) represent the speech signal and its corresponding ILPR signal for genuine and replayed signals, respectively.

where, $c(k)$, $k = 0, 1, 2, ...N-1$ are DCT coefficients, N is the number of DCT coefficients. Further, a fixed number of lower-order DCT coefficients are considered and the resultant feature vector is referred to as CILPR. It provides a compact representation of the ILPR signal between two GCIs and captures the dynamic characteristics of the glottal source signal.

To illustrate the compaction property of DCT and to show that the lower order DCT coefficients encapsulate the information contained in the ILPR, the ILPR of two pitch periods and the corresponding non-truncated CILPR for two different speech segments are depicted in Fig 4.3. From the figure, it can be noticed that the CILPR is a compressed version of the ILPR and that most of the information is contained in the first few coefficients.

4.4 Compressed Source Signal-based RAD System: Experimental Setup

In this section, the process of development of the proposed RAD system using the CILPR feature is explained. A detailed description of the experimental setup of the CILPR-based RAD system is provided.

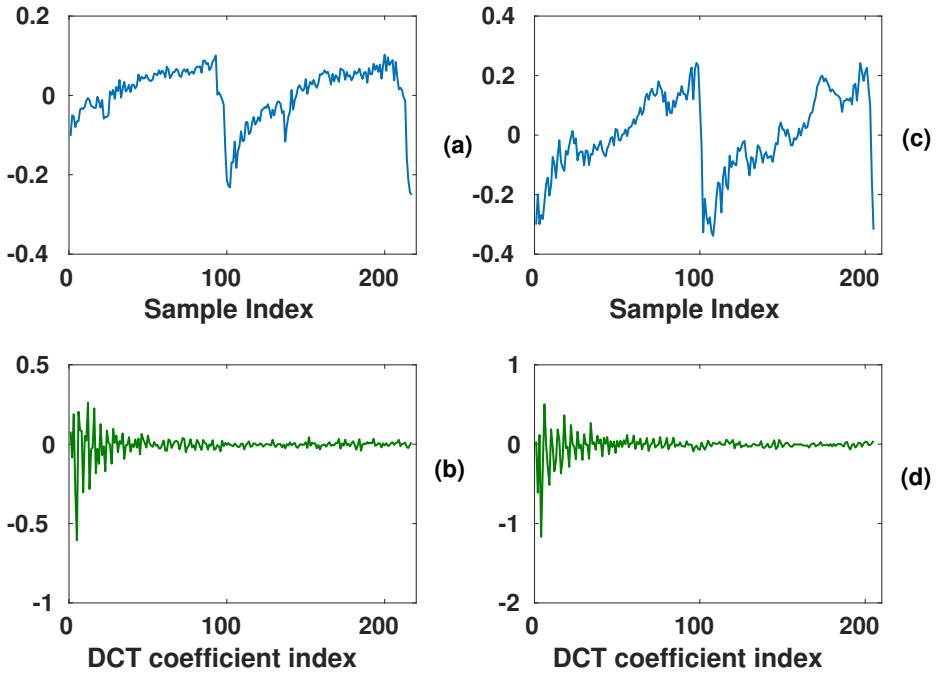

Figure 4.3: Depicting the compaction property of DCT. (a)-(b) and (c)-(d) represent the ILPR signals and their corresponding non-truncated CILPR feature for two different genuine speech segments, respectively.

CILPR feature is calculated from the ILPR in the temporal domain in a pitch synchronous manner [106]. Since 16 kHz sampling frequency is considered in this work, a 20-order LP analysis is performed. As only voiced regions are considered to compute the CILPR, a glottal activity detection algorithm is needed. Zero-frequency filtering (ZFF) based method is applied to detect the glottal activity and to estimate the GCI locations in the speech signal [97, 99]. In this case, initially, the differentiated version of the speech signal is passed through a zero frequency resonator (ZFR), and the output of the ZFR is exponentially growing or decaying in nature depending on the signal polarity. The trend of the exponential signal is removed by a moving average filter of size approximately two pitch periods and the resultant signal is referred to as the zero frequency filtered (ZFF) signal. The positive to negative zero crossings are considered as estimated GCIs of the signal. At every GCI, the positive to negative slope is termed as strength of excitation (SoE). The SoE is used to detect the glottal activity regions of the genuine and spoofed signals. The detected glottal activity regions are further used to extract the CILPR features pitch synchronously. At each GCI, the ILPR segment from the current GCI to the next GCI is considered and normalized by the norm of the segment before applying DCT-II computation. The DCT-II of the pitch synchronous

Table 4.1: Tuning the dimensionality of the CILPR feature on the development set of the ASVSpoof 2017 v2.0 database.

CILPR dimensionality	4	8	12	16	20	24	28	32	36
EER (%)	31.06	25.55	21.52	20.34	20.0	**19.68**	20.11	19.95	20.06

ILPR segment is computed for both the genuine and spoofed signals and the first few coefficients are considered, excluding the zeroth coefficient. The feature vector formed by taking lower-order DCT coefficients is referred to as CILPR in this work. Initially, an experiment is performed to obtain the optimum number of DCT coefficients to develop the spoof detection system. The lower order DCT coefficients are varied from 4 to 36 with an increment of 4 to perform the experiment. GMM-based models of 512 mixtures are built from the training subset of the database using each extracted CILPR feature. The testing is done on the development data and the results are shown in Table 4.1. From the table, it can be observed that the best performance in terms of EER is obtained with the 24- dimensional CILPR feature. After tuning the parameters of the CILPR on the development set, the system is tested on the evaluation set of the ASVSpoof 2017 v2.0 database.

4.5 Experimental Results and Discussion

The performances of the systems developed on the ASVSpoof 2017 v2.0 database using the proposed CILPR features and the baseline CQCC features are presented in Table 4.2 in terms of EER and minimum detection cost function (min. DCF). For the experiments on the development set, the training set is used to learn the GMMs of the two classes. The experiments on the evaluation set are conducted with two different training configurations. The first configuration uses only the training set to learn the GMMs while in the second configuration, the GMMs are trained on pooled training and development sets. These two configurations are referred to as C1 and C2, respectively. From Table 4.2, it can be seen that the baseline system gives an EER of 9.19% for the development set. EERs for C1 and C2 configurations of the evaluation set are 13.84% and 12.58%, respectively. The proposed system results in an EER of 19.68% for the development set. Its EER for C1 configuration is 20.66% and 15.76% for C2. The baseline and the proposed system are then fused at the score

Table 4.2: Performance comparison of different RAD systems and their score-level fusion

System	Development		Evaluation			
	Train		Train (C1)		Train + Development (C2)	
	EER (%)	min DCF	EER (%)	min. DCF	EER (%)	min DCF
Baseline: CQCC	9.19	0.455	13.84	0.739	12.58	0.662
Proposed: CILPR	19.68	0.799	20.66	0.947	15.76	0.847
Contrast: PSRMS	33.38	0.952	28.16	0.996	27.81	0.991
Fusion: CQCC + CILPR	**5.89**	**0.338**	**9.77**	**0.552**	**9.41**	**0.474**

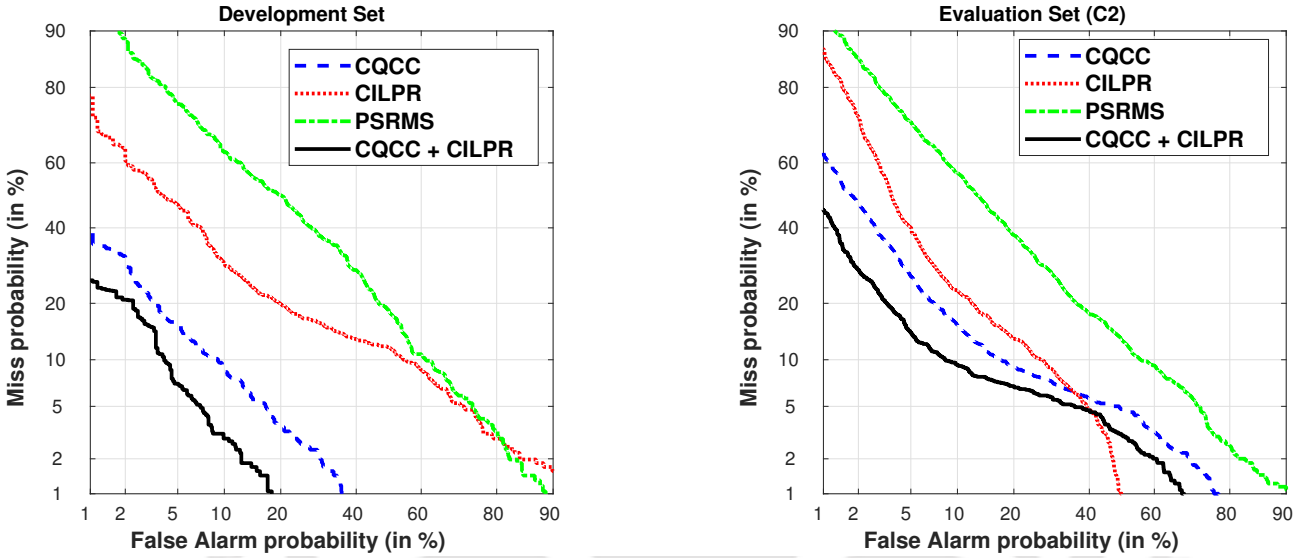

Figure 4.4: DET curves for the RAD systems developed using different kinds of features and their fusion. These curves are plotted for the development set and configuration C2 of the evaluation set

level. The fused system results in the best EER of 9.41% and minimum DCF of 0.474 for the C2 configuration of the evaluation set. For the development set, the fused system produces an EER of 5.89% and a minimum DCF of 0.338. This proves that combining source and acoustic features for replay attack detection can lead to significant performance enhancement. The detection error trade-off (DET) curves for the different spoof detection systems are given in Figure 4.4 which show similar performance trends. For contrast purposes, a system is developed using the PSRMS source feature proposed in Chapter 3. For calculating the PSRMS, first, the LP residual is estimated from the speech signal. From the LP residual, a smoothed Hilbert envelope is obtained. The peaks in the Hilbert envelope correspond to the GCI locations. The side lobes around each peak are considered to measure the peak-to-sidelobe ratios. The mean and the skewness of these ratios form a 2-dimensional PSRMS feature. GMM having 16 Gaussian components is learned from these

features. From Table 4.2, it can be noted that the PSRMS-based contrast system produces an EER of 33.38% for the development set which is about double that for the CILPR system. The contrast system results in an EER of 28.16% and 27.81% on the evaluation set for configurations C1 and C2, respectively.

Another experiment is conducted to support our hypothesis that the information extracted between two GCIs is more useful than that obtained around a GCI for replay attack detection. In this experiment, all the genuine and spoofed signals in the development set are considered. To find the separation between the two classes in the case of CILPR and PSRMS features, the Bhattacharya distances have been computed and are shown in Figure 4.5. It is observed that the Bhattacharya distance for CILPR is significantly greater than that of PSRMS which confirms our hypothesis.

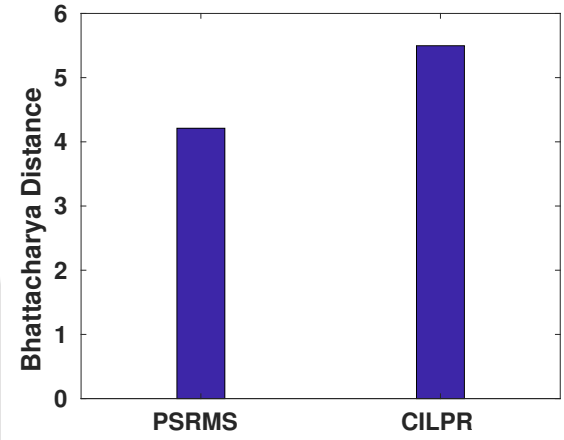

Figure 4.5: Bhattacharya distance between the genuine and replayed classes for PSRMS and CILPR features. This supports the enhanced detectability achieved with the proposed CILPR feature.

4.6 Ablation Studies

The proposed CILPR source feature is noted to yield quite competitive yet degraded replay attack detection performance compared to that of the widely used CQCC feature. Thus establishing that the analysis of the source component of speech signal provides a viable alternative to replay attack detection. In order to further improve the proposed source features, we focus on differences in the process of extraction vis-a-vis the conventional features such as LFCCs and CQCCs. It is worth highlighting that the proposed source features are computed pitch synchronously. Thus, they require the estimation of the glottal closure instants (GCI). The estimation of the GCI marks is not

only computationally intensive but also its reliability is subject to the presence of background noises in the signal. Additionally, only voiced regions are utilized in the proposed source feature, unlike the entire signal as in the case of conventional features used for replay attack detection. Since we have exploited the pitch-synchronous processing, it is obvious that such a feature cannot be computed for unvoiced/silence regions in the signal. This may result in a possible loss of information which contributes to the degraded RAD performance achieved with the proposed source features. Motivated by these observations, we seek to find out the significance of voiced and unvoiced regions of a speech signal on the RAD performance. In this section, we perform some ablation studies to try and understand the relevance of voice and unvoiced regions of speech in the detection of replay attacks. In order to do this, we perform an experiment on the ASVSpoof 2017 v2.0 database. For the same, using the method explained in Section A, the voiced and unvoiced regions are marked in training, development, and evaluation data. The relative portion of the voiced and unvoiced regions in the three partitions are shown in Figure 4.6. It can be noted that in each partition roughly about 50% of speech data corresponds to voiced and unvoiced regions. The unvoiced regions include silences and short pauses. The experiment is performed using CQCC features. The CQCC feature

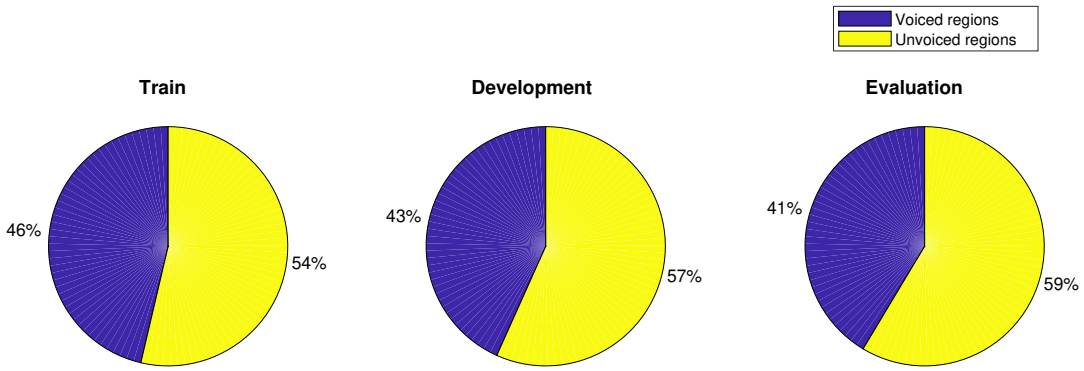

Figure 4.6: Pie chart showing the percentage of voiced vs unvoiced frames in the training, development, and evaluation sets of the ASVSpoof 2017 v2.0 database

includes 29 static coefficients plus the log-energy coefficient to which delta (Δ) and delta-delta ($\Delta\Delta$) features are appended making it a 90 dimensional feature vector. Cepstral mean-variance normalization (CMVN) is also applied to these features. From these features, two GMMs of 512

Table 4.3: Comparison of RAD systems developed with the CQCC feature using voiced and unvoiced regions of speech on the ASVSpoof 2017 v2.0 database

Conditioning applied in training and testing	Development		Evaluation	
	EER (%)	min-DCF	EER (%)	min-DCF
Voiced regions	29.18	0.986	38.3	1
Unvoiced regions	26.96	0.981	36.93	1
Entire speech	9.19	0.455	13.84	0.739

Gaussian mixture components each are learned for the genuine and spoof classes. Two systems are developed using this experimental setup. The first system is trained and tested using only glottal regions and the second system on only non-glottal regions. The results of this experiment are shown in Table 4.3. It is interesting to note that under CQCC and GMM-based replay attack detection framework, conditioning to unvoiced regions leads to better detection performance than conditioning to voiced regions. This trend is attributed to a relatively better distortion-to-speech ratio in the unvoiced region than in the voiced regions.

Both voiced and unvoiced regions are affected by replay attacks. However, the voiced regions are high-energy and relatively less impacted by these attacks. On the other hand, the unvoiced regions, which include silence regions, are low-energy regions, and the impact of replay attacks is more pronounced in these unvoiced and silence regions. Hence, it is mentioned that in the replayed speech, the distortion-to-speech ratio will be higher in the unvoiced regions compared to the voiced regions.

4.7 Conclusion

This chapter proposes the use of the ILPR signal as the speech source representation for the RAD task. The ILPR signal captures the dynamic characteristics of a signal between two GCIs and hence it has the potential to be more sensitive to the differences in genuine and replayed signals. However, the ILPR feature is calculated pitch synchronously, thus it does not produce fixed dimensional representation. On applying pitch synchronous DCT to the ILPR, a fixed dimensional feature is derived and referred to as the CILPR feature. A RAD system is developed using this feature and is evaluated on the ASVSpoof 2017 v2.0 database. It is observed that the CILPR features

provide significantly better performance than the handcrafted features. However, compared to the baseline CQCC and LFCC features, the CILPR features result in poorer performance. This degraded performance is due to the use of only the voiced regions of the source signal for extracting the CILPR feature. As explained earlier, CQCC is extracted from both voiced and unvoiced regions. Thus, to improve the performance of RAD systems, some modifications are made to the processing of the ILPR signal. These modifications and the resulting feature are discussed in Chapter 5.

5

Non-Pitch Synchronous Source Features

Contents

5.1 Introduction

In previous chapters (Chapter 3 and Chapter 4), we have developed three source-based features for replay attack detection using pitch-synchronous analysis. Through evaluation studies presented in Chapter 4, it was highlighted that there exists significant information in both voiced and unvoiced regions of the genuine and spoofed speech signals for their discrimination. It was also noted that the replay attack detection system developed using only unvoiced regions provides better performance than a similar system developed on only voiced regions. Obviously, in order to exploit the entire speech signal, we need to avoid pitch-synchronous analysis of the source signal to derive the features for replay attack detection. Motivated by that, in this chapter, we present our attempts to develop non-pitch synchronous features using frame-based processing. First, we describe two naive approaches for frame-based processing of the source (ILPR) signal to derive the feature representation. One processes the ILPR signal frames in the time domain while the other processes them in the frequency domain. Interestingly, these naive frame-based features are found to yield significantly degraded EER when compared to the earlier proposed pitch-synchronous CILPR feature on the ASVSpoof 2017 v2.0 data. To understand the possible reasons, we present a detailed analysis that motivated us to propose another novel source feature exploiting the processing of spectro-temporal patches of the source (ILPR) signal. The proposed CILPR feature is again evaluated on the ASVSpoof 2017 v2.0 and ASVSpoof 2019 databases and found to yield high improvement in terms of %EER when compared to the features extracted naively. In addition, we have evaluated the proposed CILPR feature by extracting it for voiced regions only so that we can compare it directly to pitch-synchronous features discussed earlier.

Following that, an analysis of the outcomes of the naive frame-based features is performed which motivated us towards the spectro-temporal processing of the source (ILPR) signal. The spectro-temporal processing is done in a frame-wise manner. Hence, the extraction of this feature does not require any pitch marking as a result of which one can exploit both the voiced and unvoiced regions to achieve improved discrimination.

The remainder of the chapter is organized as follows. Section 5.2 explains the non-pitch synchronous processing of the ILPR signal. In Section 5.3 the spectro-temporal processing of the ILPR signal is discussed. The proposed 2D-ILRCC features are explained in Section 5.4. The experimen-

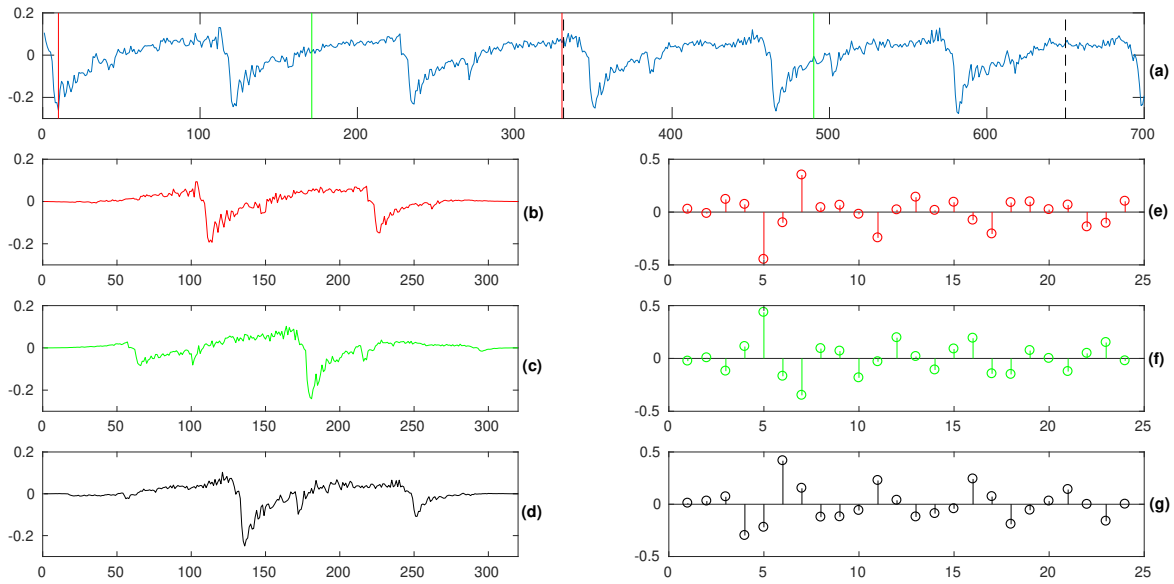

Figure 5.1: Illustrating the impact of non-pitch synchronous processing of a genuine voiced source (ILPR) signal on the CILPR feature. The analyzed source segment is shown in (a), three consecutive Hanning windowed frames extracted from the source segment are shown in (b)-(d), and their corresponding CILPR feature are shown in (e)-(g).

tal setup of the RAD system developed using the proposed features is given in Section 5.5. Finally, the conclusions are drawn in Section 5.6

5.2 Insights into Non-Pitch Synchronous Processing of the Source Signal

Motivated by the ablation study presented in Section 4.6, we make an attempt to process the source signal non-pitch synchronously. This will enable us to utilize both voiced and unvoiced regions of the speech signals in detecting the replay attack. For the same, we analyze the source signal using a fixed-length window to derive a compact representation in the form of CILPR features. This naive manner of processing can be problematic particularly for the voiced regions due to the non-synchronicity of the analysis window to the pitch cycles in a majority of the frames. As a result of that, the trends of CILPR features corresponding to homogeneous voiced regions would exhibit significant mismatch. To demonstrate this mismatch, we take a voiced source (ILPR) segment corresponding to genuine speech from the ASVSpoof 2017 v2.0 database. This voiced segment is

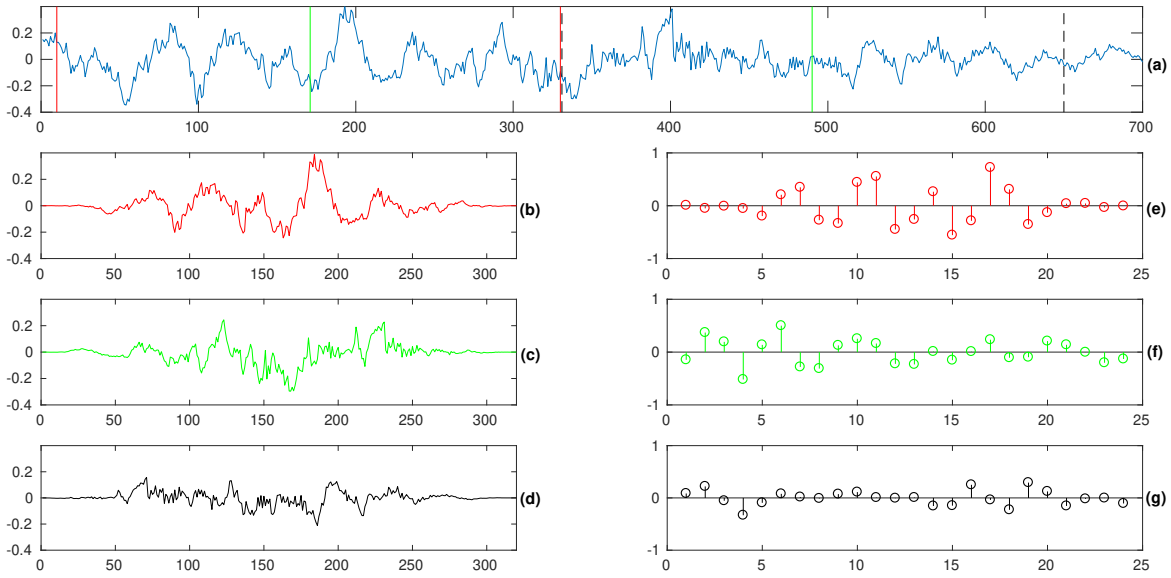

Figure 5.2: Depicting the impact of non-pitch synchronous processing of a genuine unvoiced source (ILPR) signal on the CILPR feature. The analyzed source segment is shown in (a), three consecutive Hanning windowed frames extracted from the source segment are shown in (b)-(d), and their corresponding CILPR feature are shown in (e)-(g).

analyzed using a Hanning window of size 20 ms (320 samples) with a shift of 10 ms (160 samples) and the corresponding CILPR features are computed. Figure 5.1 shows the three consecutive frames of that analysis of the source (ILPR) signal along with their corresponding 24-dimensional (C_1-C_{25}) CILPR features. From the figure, it can be noted that the CILPR features show varying trends despite the frames corresponding to the source signal having near-identical pitch periods. The main cause of this deviation in the features lies in the analysis window being non-pitch synchronous. It would be interesting to check what would be the impact of non-pitch synchronous processing in unvoiced regions of the source signal. A similar analysis was carried out on a genuine unvoiced source (ILPR) segment taken from ASVSpoof 2017 v2.0 data. Figure 5.2 shows three consecutive unvoiced ILPR frames and their corresponding CILPR features. It can be seen from this figure that the CILPR features of corresponding unvoiced frames are found to exhibit more deviated trends than those observed for the consecutive voiced frames. This is attributed to the fact that unlike the voiced regions the ILPR signal does not exhibit any defined structure.

Referring back to Figure 5.1, we can notice that on account of the non-pitch synchronous pro-

cessing, the windowed voiced signals would turn out to be approximately shifted versions of each other. Therefore, instead of processing in the time domain, if the source signal is transformed into the spectral domain the effect of any temporal shift can be mitigated to a greater extent. Motivated by this observation, we compute the spectrum of the windowed source frames using a discrete Fourier transform. Following that we compute the log-compressed magnitude spectrum and then take its DCT to derive the cepstral representation. This new feature is referred to as Integrated Linear prediction Residual Cepstral Coefficient (ILRCC). Figure 5.3 shows the log-magnitude spectrum of three consecutive windowed voiced frames considered in Figure 5.1 and their corresponding 24-dimensional ILRCC features. From the figure, it can be noticed that the features of consecutive frames show a highly similar trend unlike that observed in Figures 5.1.

In the case of non-pitch synchronous processing of unvoiced source signal, one cannot argue that the consecutive windowed frames would be shifted versions of each other as there is no periodicity in such regions. Still, we hypothesize that due to the computation of spectral magnitude, the coherence among the resulting ILRCC features would be enhanced. Figure 5.4 shows the log-magnitude spectrum for three consecutive windowed unvoiced frames considered in Figure 5.2 and their corresponding 24-dimensional ILRCC features. It can be noticed that like the voiced case, the features exhibit highly similar trends.

5.2.1 Experimental Evaluation

In this section, we experimentally evaluate the RAD performance on two kinds of features derived using non-pith synchronous processing of the source signal. One of the features is CILPR which is obtained through temporal domain processing and the other feature is ILRCC which is obtained through spectral domain processing. The RAD system employs a GMM classifier as described in Section 2.4.3. The RAD system is trained and evaluated on ASVSpoof 2017 v2 data. Table 5.1 shows the RAD performances in terms of %EER and min-DCF for two kinds of features along with their breakup on the basis of considering voiced- and unvoiced-region frames only. It can be seen from the table that ILRCC features have resulted in improved RAD performances when compared to CILPR features on both the development and the evaluation sets of the ASVSpoof 2017 v2.0 database. Further, on considering the break up of the performance in terms of voiced-

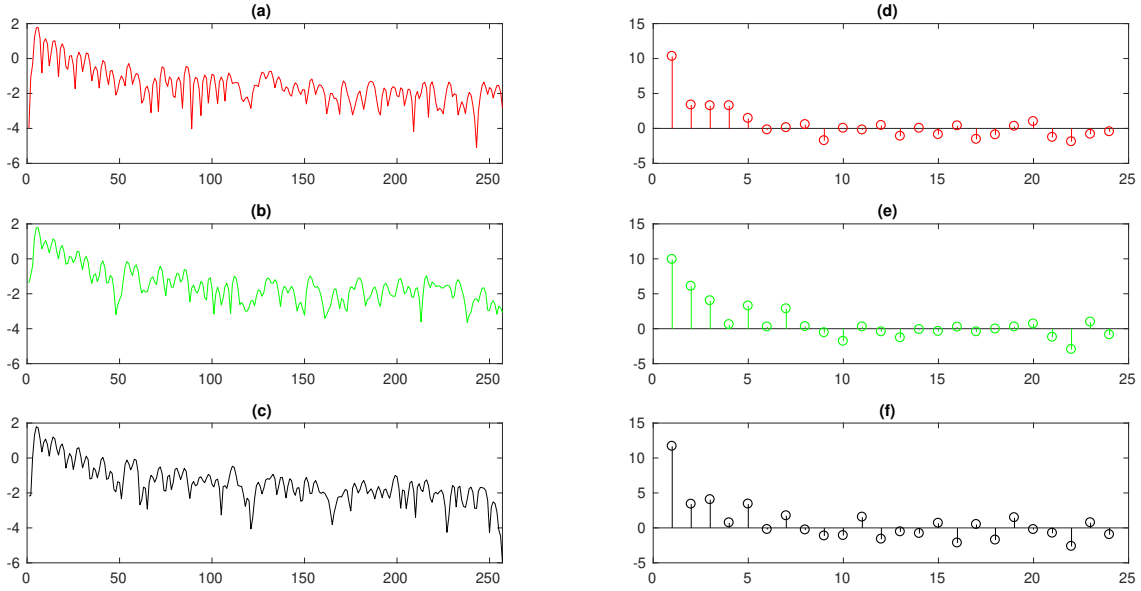

Figure 5.3: Effect of spectral domain transformation for non-pitch synchronous processing of the genuine voiced source (ILPR) signal. (a)-(c) shows the log-magnitude spectrum of the three consecutive voiced frames considered in Figure 5.1, (d)-(f) shows the corresponding ILRCC feature obtained via DCT.

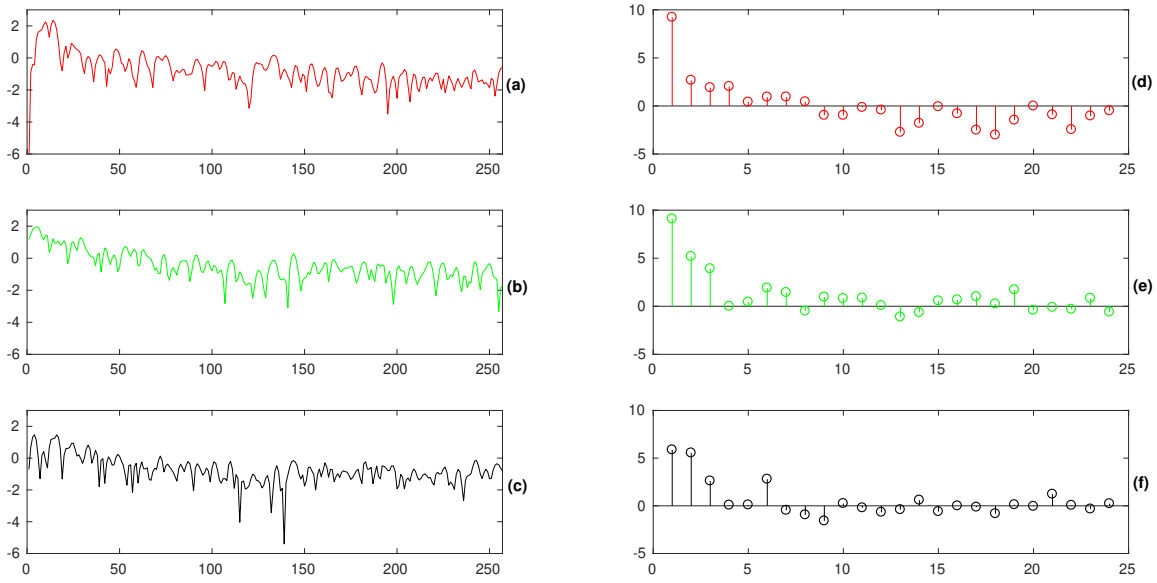

Figure 5.4: Effect of spectral domain transformation for non-pitch synchronous processing of the genuine unvoiced source (ILPR) signal. (a)-(c) shows the log-magnitude spectrum of the three consecutive unvoiced frames considered in Figure 5.2, (d)-(f) shows the corresponding ILRCC feature obtained via DCT.

and unvoiced-region frames a similar trend is noticed. This validates our hypothesis.

Table 5.1: Results of RAD systems developed using two types of non-pitch synchronous features on ASVSpoof 2017 v2.0 database. We have also computed the breakup of each performance in terms of considering only voiced-region (V) and unvoiced-region (UV) frames of the test data and those are given in the braces.

Feature	Development		Evaluation	
	EER % (V/UV)	min-DCF (V/UV)	EER % (V/UV)	min-DCF (V/UV)
CILPR	33.80 (37.24/42.06)	0.972 (0.998/1)	28.02 (30.42/37.53)	1 (1/1)
ILRCC	19.10 (25.68/34.36)	0.792 (1/1)	26.41 (33.89/38.22)	0.981 (1/1)

5.3 Capturing Source Characteristics in Spectro-Temporal Domain

In Section 5.2, we demonstrated that the spectral-domain processing of the source signal is more effective than the temporal-domain one in producing non-pitch synchronous features for replay attacks. Yet, both these features yield significantly degraded RAD performances when compared to those of pitch synchronous processing of the source signal. This motivated us to look for ways and means to improve the non-pitch synchronous processing of the source signal so that the produced features compare well with the existing frame-based features such as LFCCs and CQCCs.

It is worth highlighting that the spectral domain processing of the source signal is found to enhance the coherence among the spectral representations for consecutive frames corresponding to both voiced and unvoiced regions. As a result of that the neighbouring frames are expected to exhibit highly similar spectral structures which can be exploited for more effective capture of the source signal dynamics. To illustrate this, we plot the spectrograms for a pair of genuine and replayed speech examples taken from the ASVSpoof 2017 v2.0 database as shown in Figure 5.5. For the sake of argument, let us consider an overlaid temporal slice covering a set of neighboring frames in both spectrograms as shown in the figure. On comparing the underlying regions of the genuine and replayed spectrograms, we can notice that there exists enhanced temporal continuity of the underlying spectral structure for the former rather than the latter. It is hypothesized that by capturing the same, better discrimination between genuine and replayed signals can be achieved. This motivated us to develop a novel feature for RAD as described in the following section.

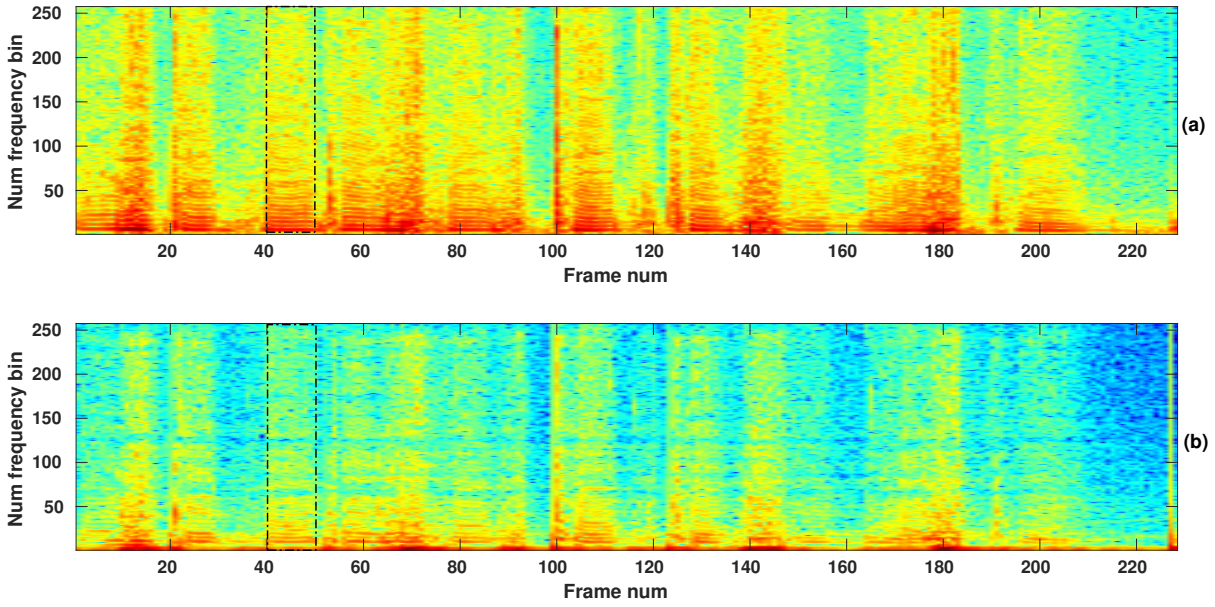

Figure 5.5: Spectrograms of ILPR signals corresponding to a (a) genuine and (b) replayed speech signal pair taken from the ASVSpoof 2017 v2.0 database. The dotted rectangles show spectro-temporal patches which allow us to capture the temporal continuity of spectral structure if any. This in turn may enable better detection of replay attacks.

5.4 Spectro-Temporally Compressed ILPR Feature

In this section, we describe the derivation of a novel frame-based feature obtained by compressing a patch of the time-frequency representation of ILPR to better capture the spectro-temporal modulations in the source signals. For compressing the time-frequency patch of the source signal, we have used a 2-dimensional discrete cosine transform (2D-DCT). The effectiveness of joint spectro-temporal features based on the 2D-DCT has been demonstrated for automatic speech recognition (ASR) application [107], detection of place of articulation in unvoiced stops [108], and evaluation of pathological speech [109,110]. For real-time applications, we have considered a causal time-frequency block in extracting the proposed frame-based features. First, the speech signal is short-time analyzed with a Hanning window of size 20 ms (320 samples) and a shift of 5 (80 samples). For each frame, the inverse filtering is performed to yield the residual signal using 20-order linear prediction analysis. The resulting source (ILPR) signal frame is processed using a 512-point fast Fourier transform to yield logarithmically compressed magnitude spectra. Following that, for each analysis

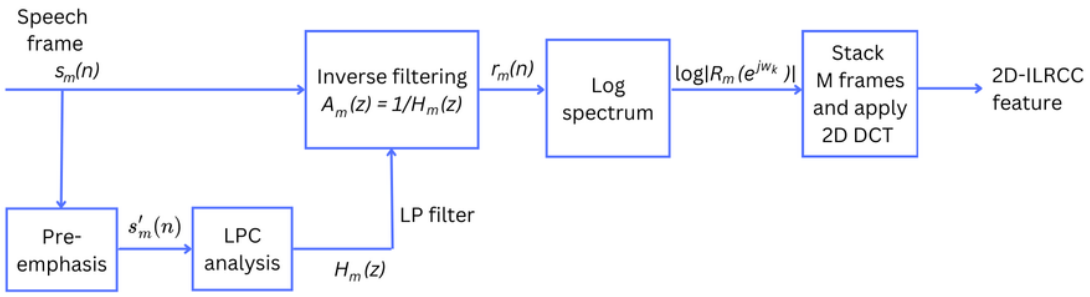

Figure 5.6: Block diagram of the process of extraction of the 2D-ILRCC feature

frame, a spectro-temporal 2D patch is created by stacking the log-magnitude spectra of past $(M-1)$ frames. So created 2D patch is projected to a 2D cosine basis. Let P be a spectro-temporal 2D patch of size $N \times M$, where N and M represent the spectral and temporal extents of the 2D patch, respectively. The bottom left entry of the patch corresponds to the start of time and frequency indices. Then, the kth row and lth column entry of the 2D-DCT projection C of the patch P is given by,

$$C_{kl} = \frac{2\omega_k\omega_l}{\sqrt{NM}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_{nm} \cos\left\{\frac{\pi l(2a+1)}{2M}\right\} \cos\left\{\frac{\pi k(2b+1)}{2N}\right\}, \tag{5.1}$$

where $k = 0, 1, \ldots, N-1$, $l = 0, 1, \ldots, M-1$, and

$$\omega_k = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0, \\ 1 & \text{if } k \neq 0, \end{cases}$$

$$\omega_l = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } l = 0, \\ 1 & \text{if } l \neq 0. \end{cases}$$

The horizontal DCT coefficients of matrix C correspond to the temporal characteristics, whereas the vertical coefficients correspond to the spectral envelope. Later, we consider low-order 2D-DCT coefficients, which provide a compact representation of the spectro-temporal modulations contained in the 2D patches.

In this case, the spectral extent N is the number of frequency bins used to compute the spectrum, and the temporal extend M is the number of frames. For each overlapping patch, we apply a 2D-DCT to compute a set of DCT coefficients. Suitable truncation of the resulting DCT coefficient matrix followed by linearization is used as the feature vector and is referred to as the 2D-ILRCC

feature.

5.5 Experimental Setup and Discussion

This section first provides a detailed description of the experiments performed to tune the dimensionality of the 2D-ILRCC feature. Next, the performance evaluation of the 2D-ILRCC feature on the ASVSpoof 2017 v2.0 database is presented. Finally, the effectiveness of the 2D-ILRCC feature is validated on the ASVSpoof 2019 database.

5.5.1 Tuning of 2D-ILRCC Feature Dimensionality

The parameters of the 2D-ILRCC features are tuned on the ASVSpoof 2017 v2.0 database. The source (ILPR) signals corresponding to the genuine and replayed speech are short-term processed using Hanning windowed frames of 20 ms (320 samples) and a shift of 5 ms (80 samples). Each frame is then transformed to the spectral domain using 512-point FFT. On account of the chosen low framerate, the resulting spectrogram of a source signal exhibits high temporal continuity. Given the source spectrogram, we form a causal patch consisting of all spectral bins and its temporal dimension empirically set to 70 ms which is equivalent to the temporal context of 11 frames. This choice of temporal extent is made as a compromise between the average duration of voiced and unvoiced units of speech. Thus, the resulting patch is of size 512×11. Such patches are produced every 2 frames so that the frame rate of the final feature becomes 10 ms, i.e., the typical frame rate used in speech analysis. Each patch is then applied with 2D-DCT. The resulting 2D coefficient matrix after suitable truncation and linearisation yields the 2D-ILRCC feature vector. Mean-variance normalization is applied to the extracted features on an utterance basis. The normalized features are used to develop the RAD system consisting of two GMMs of 512 mixture components each modeling the density of genuine and replayed speech data.

A grid search is performed to tune the truncation of the 2D-DCT coefficient matrix through a rectangular lifter. Initially, the horizontal dimension of the rectangular lifter is set to the maximum possible value of 11. While keeping the horizontal dimension of the lifter constant, its vertical dimension is varied from 20 to 120 in steps of 20. It is observed that the best performance in terms of EER is achieved when the vertical dimension of the rectangular lifter is set to 40 as shown in

Table 5.2: Tuning the dimensionality of the proposed 2D-ILRCC feature for RAD on the ASVSpoof 2017 v2.0 database

Lifter Dimension		EER (%)	
Vertical	**Horizontal**	**Development**	**Evaluation**
10	4	21.84	18.44
10	5	20.98	17.40
10	6	20.27	16.60
20	4	16.44	14.95
20	5	16.08	14.84
20	6	16.14	13.75
30	4	14.07	12.21
30	5	13.06	11.93
30	6	12.83	11.96
40	4	12.73	11.03
40	5	11.98	11.33
40	6	**11.90**	**10.87**
40	7	12.89	10.99

Table 5.2. Subsequently, a finer grid search is performed with the vertical dimension of the lifter being varied from 10 to 40 in steps of 10 and the horizontal dimension of the lifter being varied from 4 to 7 in steps of 1. From Table 5.2 it can be seen that the best performances are obtained for both the development and evaluation sets when the rectangular lifter is of size 40×6.

5.5.2 Performance on ASVSpoof 2017 v2.0 Database

The performance of the RAD system developed using the proposed 2D-ILRCC features and its comparison to the baseline systems developed with CQCC and LFCC features on the ASVSpoof 2017 v2.0 database are shown in Table 5.3. It can be seen from the table that the 2D-ILRCC features result in EERs of 11.90% and 10.87% for the development and evaluation sets, respectively. Compared to the baseline features, the proposed feature achieves improved performance on the evaluation set of the considered database. This shows the effectiveness of the 2D-ILRCC features for RAD. It would be interesting to compare the RAD performance of the proposed non-pitch synchronous feature with that of the pitch-synchronous CILPR feature described in Chapter 4. On comparing the RAD performances in Table 4.2 and Table 5.3, it is observed that the proposed non-pitch synchronous 2D-ILRCC features achieve a relative improvement of 39.53% and 47.38%

in terms of EER on the development and evaluation sets of the ASVSpoof 2017 v2.0 database respectively. This improvement is attributed to a novel way of addressing the distortion caused by non-pitch synchronous processing of the source signal by enhancing the coherence between the fixed-length frames by applying the proposed spectro-temporal processing.

Table 5.3: Performance comparison of CQCC, LFCC, and 2D-ILRCC features on the development and evaluation sets of the ASVSpoof 2017 v2.0 database.

	EER (%)	
Feature	**Development**	**Evaluation**
LFCC	17.11	16.89
CQCC	9.19	13.84
2D-ILRCC	11.90	10.87

5.5.3 Validation on ASVSpoof 2019 Database

The performance of the 2D-ILRCC features is also validated on the PA set of the ASVSpoof 2019 database. The PA set contains three partitions similar to the ASVSpoof 2017 v2.0 database. These are the train, development, and evaluation partitions. In order to validate RAD performance on the ASVSpoof 2019 database, first the 2D-ILRCC features are extracted from the train partition and then a RAD system is developed consisting of two GMMs of 512 components each modeling the density of genuine and replayed speech data. The RAD performance is then assessed on the development and evaluation sets in terms of EER. Another metric called the tandem-DCF (t-DCF) is introduced to measure the joint performance of RAD and automatic speaker verification (ASV) systems. This metric is used to determine the affect of replay attacks on an ASV system and how effective the feature is in negating this impact. The ASV system in this context is developed using DNN-based x-vector speaker embeddings with a probabilistic linear discriminant analysis (PLDA) back-end. Table 5.4 shows the results of the proposed 2D-ILRCC feature as well as the baseline CQCC and LFCC features on the PA subset of the ASVSpoof 2019 database. From the table, it can be observed that the RAD system designed using the 2D-ILRCC features achieves comparable results with regard to the baseline systems in terms of both EER and t-DCF.

Table 5.4: Performance comparison of different features on the development and evaluation sets corresponding to the PA portion of the ASVSpoof 2019 database

Feature	Dev		Eval	
	t-DCF	EER (%)	t-DCF	EER(%)
LFCC	0.214	9.89	0.260	11.44
CQCC	0.207	10.39	0.261	11.66
2D-ILRCC	0.199	9.15	0.272	11.35

5.6 Conclusion

This chapter introduces the non-pitch-synchronous processing of the source (ILPR) signal. This kind of processing enables us to extract source features from both the voiced and unvoiced regions of the speech signal. Then a spectro-temporal processing of the source signal is proposed which results in the extraction of the 2D-ILRCC feature. A RAD system is developed using this new feature and a GMM classifier on the ASVSpoof 2017 v2.0 database. This system outperforms the CQCC and LFCC-based RAD systems. Next, a RAD system is also developed with the 2D-ILRCC features and GMM back-end on the ASVSpoof 2019 database. On this database, the 2D-ILRCC achieves competitive performance when compared to the baseline features CQCC and LFCC. This proves that the 2D-ILRCC feature generalizes well to both the databases and provides RAD performance which is as good as the baseline systems. However, the proposed features are still unable to outperform the baseline features. This is due to the fact the proposed features are extracted only from the source component of the speech signal and the filter component is discarded. Hence in Chapter 6, a new feature for RAD is proposed that makes use of both the source and filter components of speech.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

6

Combination of Source and Filter Features

Contents

6.1 Introduction

In the preceding chapter (Chapter 5), we presented studies showing that there exists significant information in the source component of the speech signal for RAD. With the help of only the source component, we were able to achieve RAD performance that is at par with the CQCC-based RAD system. However, we know that replay attack artifacts impact both the source and filter components, and therefore spoofing information is present in both these components. The CQCC features happen to capture the information present in both the source and filter components of the speech signal. This is attributed to the fine sampling of the low-frequency region due to the utilization of the CQT-transform. Thus, it is expected that if the 2D-ILRCC features proposed in the previous chapter are enhanced with spoofing information available in the filter component, the resulting combined features may be able to attain state-of-the-art RAD performance. To pursue that objective, in this chapter, we attempt to extract a novel feature in a non-pitch synchronous manner that can simultaneously incorporate the spoofing information available in both the source and filter components of the speech signal. We have already hypothesized that the channel noise having differing spectro-temporal signatures would be better captured if the speech signal is decomposed into source and filter components. Motivated by that, we propose a novel feature set for replay attack detection by capturing long-term trends in the source as well as filter components of the input speech. For this purpose, we employ a moderate-order linear prediction (LP) analysis of pre-emphasized input speech frames to achieve an effective source and filter separation. Following that, the spectral representations are obtained for both components, which are then stacked temporally to capture long-term trends.

The proposed features are first evaluated on the ASVSpoof 2017 v2.0 database in the context of a RAD system with a GMM back-end. Later, using the proposed features as the front-end, a deep residual network (ResNet-18) is employed as a back-end to develop a RAD system. This RAD system is evaluated on the PA set of the ASVSpoof 2019 database. Next, a minor modification is made to the proposed features to obtain gram-based features. These features are then used to develop a RAD system with a ResNet-18 back-end. The gram-based RAD system is also evaluated on the PA set of the ASVSpoof 2019 database and contrasted with CQT-based gram features.

The rest of the chapter is organized as follows. The proposed features are explained in Section 6.2.

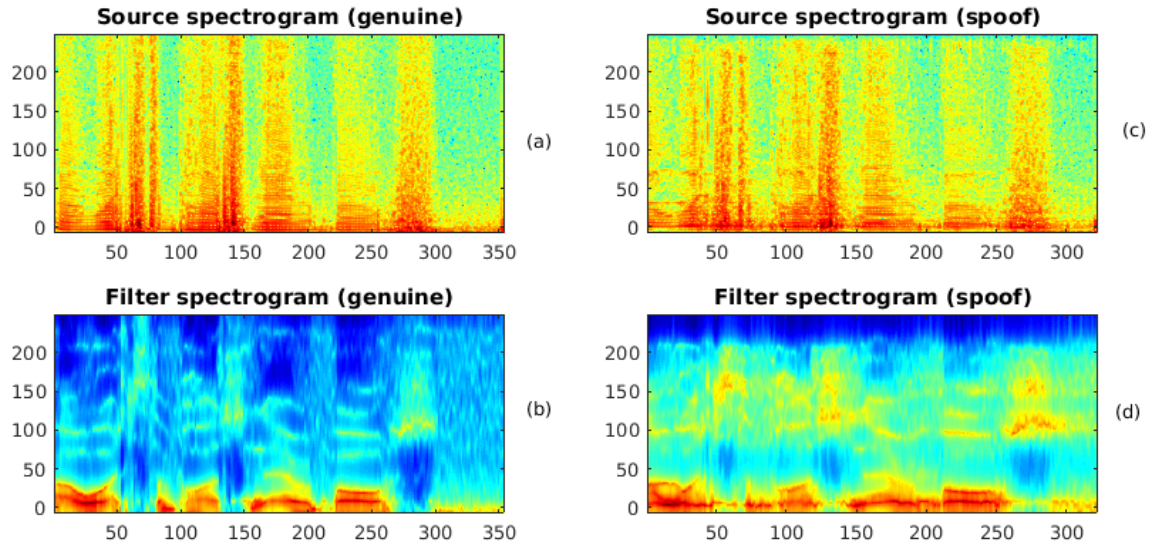

Figure 6.1: Spectrograms of source and filter components of genuine and replayed speech signals. (a) Source spectrogram of genuine signal (b) LPC spectrogram of genuine signal (c) Source spectrogram of replayed signal (d) LPC spectrogram of replayed signal.

The experiments conducted using the proposed features and the results obtained for the GMM back-end-based RAD system are described in Section 6.3. Next, in Section 6.4, we describe the development of a deep learning-based RAD system, in particular, the ResNet-18 model and the corresponding RAD performances evaluated on the ASVSpoof 2019 database. In Section 6.5, the combined source-filter time-frequency features are introduced and the performances of RAD systems developed using these features are reported. Section 6.6 provides a comparison of the proposed features with contemporary features used in other works. Finally, the conclusions drawn from this study are summarized in Section 6.7.

6.2 Computation of Combined Source and Filter Features

This section presents the mathematical steps involved in computing another novel RAD feature. These features are derived by combining the source-based features discussed earlier with the source-linked filter features.

Let $s_m(n)$ denote the m^{th} frame of the input speech signal and $s'_m(n)$ denote the corresponding pre-emphasized version. In the LP analysis of each frame, the current speech sample is predicted

as the weighted sums of the past P samples as

$$\tilde{s'}_m(n) \approx \sum_{p=1}^{P} \alpha_p s'_m(n-p) \tag{6.1}$$

The prediction residual signal $r_m(n)$ is of the form

$$\begin{aligned} r_m(n) &= s_m(n) - \tilde{s'}_m(n) \\ &= s_m(n) - \sum_{p=1}^{P} \alpha_p s'_m(n-p) \end{aligned} \tag{6.2}$$

On taking z-transform of both sides of Eq. (6.2), it can be argued that the prediction residual (or source) signal is the output of analysis system function $A_m(z)$ processing the input speech as

$$A_m(z) = \frac{R_m(z)}{S_m(z)} = 1 - \sum_{p=1}^{P} \alpha_p z^{-p} \tag{6.3}$$

The synthesis system function $H_m(z)$ happens to be the inverse of $A_m(z)$ and corresponds to the vocal filter involved in speech production. The $H_m(z)$ can be expressed as

$$H_m(z) = \frac{S_m(z)}{R_m(z)} = \frac{1}{1 - \sum_{p=1}^{P} \alpha_p z^{-p}} \tag{6.4}$$

Now, the analyzed speech frame's source and filter components are transformed into the spectral domain. For deriving spectral representation of the source signal, first K-point discrete Fourier transform of $r_m(n)$ is computed and then converted to logarithmic power spectrum as

$$r_m(n) \longleftrightarrow R_m(e^{j\omega_k}) \Rightarrow \log |R_m(e^{j\omega_k})|^2 \tag{6.5}$$

The spectral representation for the filter component is derived by evaluating the filter system function $H_m(z)$ at K uniformly sampled points on the unit circle and converting it to logarithmic power spectrum as

$$H_m(z) \longrightarrow H_m(e^{j\omega_k}) \Rightarrow \log |H_m(e^{j\omega_k})|^2 \tag{6.6}$$

For capturing long-term trends, M consecutive frames are stacked together to obtain two $K \times M$ sized log-spectrograms ($\mathcal{A}$ and $\mathcal{B}$) corresponding to source and filter components as

$$\mathcal{A}_{km} = \log |R_m(e^{j\omega_k})|^2 \text{ and } \mathcal{B}_{km} = \log |H_m(e^{j\omega_k})|^2$$

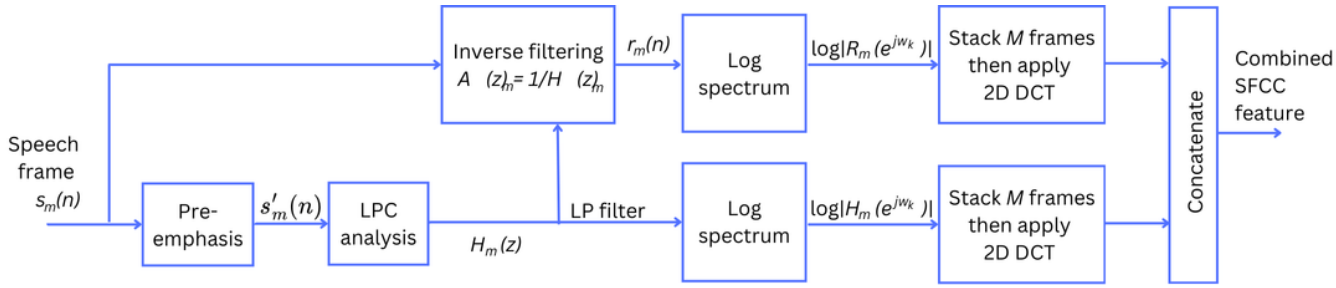

Figure 6.2: Block diagram showing the process of extracting the proposed combined source-filter cepstral coefficient (CSFCC) feature corresponding to a speech frame.

where $0 \leq k \leq K-1$, $0 \leq m \leq M-1$.

Following that, each log-spectrogram is applied with 2-D DCT to derive corresponding cepstral representations as

$$\mathcal{C}_{pq} = \eta_p \eta_q \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \mathcal{A}_{km} \cos \frac{\pi(2k+1)p}{2K} \cos \frac{\pi(2m+1)q}{2M}$$

$$\mathcal{D}_{pq} = \eta_p \eta_q \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \mathcal{B}_{km} \cos \frac{\pi(2k+1)p}{2K} \cos \frac{\pi(2m+1)q}{2M}$$

where $\eta_p = \begin{cases} \frac{1}{\sqrt{K}} & p = 0 \\ \sqrt{\frac{2}{K}} & 1 \leq p \leq K-1 \end{cases}$

and $\eta_q = \begin{cases} \frac{1}{\sqrt{M}} & q = 0 \\ \sqrt{\frac{2}{M}} & 1 \leq q \leq M-1. \end{cases}$

After applying suitable truncation of the cepstral coefficients, the entries of matrices $\mathcal{C}$ and $\mathcal{D}$ are linearized and concatenated to yield a final feature representation and the same is referred to as the combined source-filter cepstral coefficients (CSFCC) in this work. The block diagram of the overall processing of the CSFCC features is shown in Figure 6.2. It is worth highlighting that the source component in the CSFCC feature is contributed by the linearized matrix $\mathcal{C}$ which is the same as the 2D-ILRCC features discussed in Section 5.4. The filter component in the CSFCC feature is contributed by the linearized matrix $\mathcal{D}$ which can also be interpreted as 2D linear prediction cepstral coefficients (2D-LPCC).

6.3 Experimental Evaluation

6.3.1 CSFCC-based RAD System

This subsection describes the process of development of a RAD system using the proposed CSFCC features. The input speech signal is first decomposed into its source and filter components. The source component is obtained by calculating the integrated linear prediction residual (ILPR) signal. The ILPR signal is then processed using short-term analysis to compute the 2D-ILRCC feature as described in Section 5.5.1. This is the final source representation of the input speech signal.

Next, to obtain the filter component the input signal is LPC analyzed using a 20-order LP filter using a short time window size of 20 ms and shift of 5 ms, and the LPC coefficients are calculated. The LPC coefficients of each frame are passed through the LP filter and then the fast-Fourier transform of 512 points is computed. The FFT coefficients for each frame are then stacked together in a matrix. Then the same processing that is applied to the source component is done. From the resultant matrix, 70 ms overlapping patches are taken and then 2D-DCT is applied to the patches. The 2D-DCT coefficients of each patch are suitably truncated and then stacked together to get the final representation for the filter component.

Now since the sizes of the overlapping patches are the same for both source and filter components, the two resultant matrices are combined on each patch, and thus the final feature vectors are generated for each patch. This combined output matrix is called the CSFCC features. The proposed features are used to develop a GMM-based RAD system as explained in Section 2.4.3.

6.3.2 Tuning the Dimensionality of the CSFCC Feature

As discussed earlier, the CSFCC features happen to be the concatenation of 2D-ILRCC and 2D-LPCC features. These components have been derived separately including the tuning of their dimensionality. The dimensionality of the source component (2D-ILRCC) is fixed as 240 as already discussed in Section 5.5.1. To tune the dimensionality of the filter component of CSFCC features, a GMM-based RAD system is developed on 2D-LPCC features on the ASVSpoof 2017 v2.0 database. The genuine and replayed signals of the database are short-term processed using a Hanning window and LP analysis is performed on the windowed frames. The windowed frames are transformed to the

spectral domain using a fast-Fourier transform. The 2D-LPCC features are then extracted and two GMMs are trained as discussed in Section 6.3.1. A grid search is performed to tune the truncation of the 2D-DCT coefficient matrix through a rectangular lifter. The vertical dimension of the lifter is varied from 10 to 40 in steps of 10. The horizontal dimension of the lifter is varied from 5 to 6 in steps of 1. The results of the tuning experiment for both the development and evaluation sets of the ASVSpoof 2017 v2.0 database are shown in Table 6.1. From the table, it can be observed that the best performance is obtained when the vertical and horizontal dimensions of the rectangular lifter are set to 30 and 5 respectively.

Table 6.1: Results of dimension tuning experiments of the filter component of the CSFCC feature on the ASVSpoof 2017 v2.0 database

2D-Lifter Dimensions		**EER (%)**	
Vertical	**Horizontal**	**Development**	**Evaluation**
10	5	18.53	17.09
20	5	17.01	16.39
20	6	17.4	16.8
30	5	**14.14**	**14.18**
40	5	16.46	15.16

6.3.3 Evaluation of the CSFCC Feature on the ASVSpoof 2019 Database

In this sub-section, the experimental evaluation of the proposed CSFCC features performed on the PA set of the ASVSpoof 2019 database is presented. From the genuine and replayed speech signals the CSFCC features are extracted as explained in Section 6.3.1. The CSFCC features are of 390 dimensions, i.e. 240 for the source component and 150 for the filter component. A GMM-based RAD system is developed using the CSFCC features on the ASVSpoof 2019 database. The performance of the RAD system in terms of EER and t-DCF is presented in Table 6.2. The CSFCC-based RAD system results in an EER of 9.85% and a t-DCF of 0.233 and outperforms the baseline CQCC system by a considerable margin. This shows the effectiveness of the proposed CSFCC features.

Table 6.2: Performance comparison of baseline and proposed RAD systems on the PA evaluation set of the ASVSpoof 2019 database

Feature	t-DCF	% EER
CQCC	0.261	11.66
CSFCC	0.233	9.85

6.4 Deep Learning Approaches for RAD

Thus far in this thesis, the major focus has been on developing novel features for RAD and the systems have been built with a relatively simple GMM back-end. This approach allows for easy benchmarking of the proposed features. However, to achieve state-of-the-art performances, it is essential to incorporate more complex machine learning models as the back-end. In recent years, deep learning-based models have been widely adopted for RAD systems, and have yielded state-of-the-art performances.

Various kinds of deep learning approaches have been explored for building RAD systems. Deep learning models have been used in the form of front-end feature extractors, back-end classifiers as well as end-to-end solutions for RAD. In one of the earliest works done in [111], bottleneck features were generated using a deep neural network for spoofing detection. Several features extracted from deep learning architectures have been explored in [112]. In [34], high-frequency cepstral coefficients were used as the front-end, and a deep neural network was used as a back-end classifier. A RAD system developed using segment-based linear filter bank features and an attention-enhanced DenseNet-bidirectional long short-term memory was proposed in [113]. End-to-end solutions for RAD have been investigated in [114, 115].

The most common deep classifiers used in replay attack detection are different variations of light convolutional neural networks (LCNN) and deep residual networks (ResNet). The 9-layer LCNN was the best system in the ASVSpoof 2017 challenge [36], where a Max-Feature-Map (MFM) activation is used after each convolution operation. It also performed well in ASVSpoof 2019 challenge [44], [116]. The ResNet variations used in ASVSpoof 2019 challenge achieved great performance in both PA and LA subtasks [116], [46], [45], [57], [117]. Motivated by the success of ResNet based architectures, we have adopted a ResNet-18 classifier as the back-end for our experiments. It is expected that the use of the ResNet-18 classifier will further enhance the performance

of the RAD systems. The details of its implementation are discussed in subsequent sections.

6.4.1 ResNet Architecture

A general understanding when solving a complex problem using deep learning is that the deeper a network is, the better it is at learning the non-linear relationships between the input and output data. This is because as the network grows deeper, the model flexibility increases with an increased number of parameters. However, it has been observed that after a certain depth, the performance starts to degrade. This phenomenon is attributed to the vanishing gradient problem. The vanishing gradient problem occurs when the network becomes too deep and the gradients shrink to zero after repeated application of the chain rule. Residual networks aim to solve this problem by introducing the concept of a residual block as shown in Figure 6.3. Residual networks were first proposed in [118]. It can be seen from the figure that there is a direct connection between some layers skipping the layers in between. This is called the skip connection and the layers together form a residual block. ResNets are constructed by stacking together several residual blocks. From Figure 6.3, we can see that input to the first weight layer is x and the output from the last layer is $F(x)$ if the skip connection did not exist. However, with the addition of the skip connection the output from the last layer becomes $F(x) + x$.

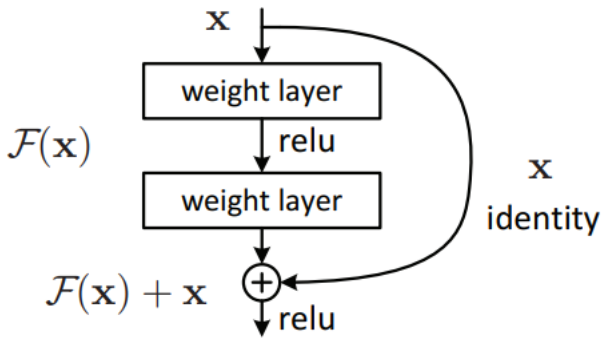

Figure 6.3: Residual block

Let us assume that the output from the last layer is $H(x)$. Mathematically, we can write:

$$H(x) = F(x) + x \tag{6.7}$$

Here, $H(x)$ is the underlying mapping to be fit by the training process. However, in ResNets the

idea is to learn the residual mapping $H(x) - x$ instead of $H(x)$. This is done by explicitly allowing the layers to approximate the residual function $F(x)$. The original function thus becomes $F(x) + x$. A problem that may arise in this case is that the dimensions of x and $F(x)$ may not be the same. This issue can be addressed by using any one of two approaches. The first approach is to apply zero padding to make the dimensions the same. The second approach is to use a convolution layer from x to the addition of $F(x)$.

ResNets help to solve the problem of vanishing gradients by providing alternate paths for the gradients to flow through via the skip connections. It also helps the connections by allowing the model to learn the identity functions which ensures that the higher layer will perform at least as good as the lower layer, and not worse. The complete idea is to make $F(x) = 0$. So, in the end, we have $H(x) = x$ as a result. This means that the value coming out from the activation function of the identity blocks is the same as the input from which we skipped the connections.

Table 6.4 shows the architecture of the different ResNet variations as mentioned in [118].

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 6.4: Different architectural variants of the ResNet used in the literature

6.4.2 Experimental Evaluation

In this section, the proposed CSFCC and the baseline features are evaluated on the PA set of the ASVSpoof 2019 database using a ResNet-18 classifier. CSFCC are extracted as explained in Section 6.3.1 and the baseline features are extracted as explained in Section 2.5.1. The extracted features are used as input to the ResNet-18 back-end. The details of the ResNet-18 architecture

used in the work and the training process are described next.

6.4.2.1 Model and Training Details

In this work, we use a variant of the ResNet-18 architecture as the back-end. We use the same four residual blocks present in the general ResNet-18. However, we use a self-attention layer after the residual blocks instead of a global average pooling layer that is employed in regular implementations of ResNet-18. The architecture used in this work has been taken from [119]. The complete details of the different layers in the model are given in Table 6.3. The first layer is the input layer, which takes the front-end features as input and passes them on to the next layer, a convolutional layer. After this, there are four residual blocks, which consist of several convolutional layers, as shown in the table. The output from the last residual block is passed on to a self-attention layer. Next, there are two fully connected layers. The output from the first fully connected layers is used as embeddings. The second fully connected layer is the output layer with two nodes, signifying the genuine and replayed likelihoods. The log-likelihood score is calculated from the output of these two nodes by first taking their logarithm and then subtracting one from the other. The required

Table 6.3: Architecture of the ResNet model used as back-end in the CSFCC-based RAD system

Layer	Kernel size	Filters	Output size
Input	-	-	$32 \times 1 \times 390 \times 750$
Conv	9x3	16	$32 \times 16 \times 39 \times 375$
BatchNorm	-	-	$32 \times 16 \times 39 \times 375$
Res1	3×3	64	$32 \times 64 \times 20 \times 188$
Res2	3×3	128	$32 \times 128 \times 10 \times 94$
Res3	3×3	256	$32 \times 256 \times 5 \times 47$
Res4	3×3	512	$32 \times 512 \times 3 \times 24$
Conv	3×3	256	$32 \times 256 \times 1 \times 12$
BatchNorm	-	-	$32 \times 256 \times 12$
Self-Attention	-	-	32×512
Fully connected	-	-	32×256
Fully connected	-	-	32×2

features (CSFCC/CQCC) are first extracted to train this ResNet-18 model. Since the speech files from which these features are calculated are of different durations, the number of frames for each file may differ. However, to use these features as input to the ResNet-18 model, they must be the same size. This similarity in size is achieved by setting the features to be of a fixed frame length

of 750 frames. Repeat padding is used to make the short trials of the desired length. The first 750 consecutive frames are considered for the longer trials, and the remaining ones are discarded. Once this is done, the model is trained with these features using a batch size of 32. The model output is a confidence score that signifies the classification result. The softmax loss function and adaptive moment estimation (Adam) optimizer are used to update the weights of the ResNet model.

The efficacy of different RAD systems using the ResNet-18 back-end is validated on the PA set of the ASVSpoof 2019 database. The CSFCC and CQCC features are extracted from this database's training subset. The dimensionality of the CSFCC feature is 390 dimensions, and CQCC features are of 90 dimensions. Two separate models are then trained using the respective features. The features of the development and evaluation sets are used to obtain the classification scores, which are then used to calculate the EER and t-DCF.

6.4.2.2 Results and Discussion

The performances of the CSFCC and CQCC-based RAD systems on the PA set of the ASVSpoof 2019 database are shown in Table 6.4. It can observed that the proposed CSFCC features outperform the baseline CQCC features by a significant margin both in terms of EER and t-DCF. In the evaluation set, the CSFCC-based RAD system results in an EER of 1.97% whereas the CQCC-based RAD system gives an EER of 4.35%. Thus the proposed RAD system results in a relative improvement of 54.71% compared to CQCC-based RAD system. This proves the advantage of the proposed features for building RAD systems.

Table 6.4: Performance comparison of baseline and proposed RAD systems on the PA evaluation set of the ASVSpoof 2019 database

Feature	t-DCF	% EER
CQCC	0.135	4.35
CSFCC	0.061	1.97

To demonstrate the ability of the ResNet-18 network to effectively model the feature space, the t-distributed stochastic neighborhood embedding (t-SNE) plots for the CSFCC and CQCC features are shown in Figure 6.5. The t-SNE plot is drawn for the CQCC and CSFCC features of the evaluation set of the PA partition of the ASVSpoof 2019 database. The features are passed through the trained ResNet-18 network, and the 256-dimensional embeddings extracted from the second to

the last fully connected layers are considered. These 256-dimensional embeddings are transformed into a 2D space for visualization using the t-SNE technique. From the plots, it can be seen that the embeddings of CSFCC features for the genuine and replayed cases are far more well-separated when compared to those of CQCC features. For this reason, the proposed CSFCC and ResNet-18-based RAD system performs significantly better than the one built using CQCC features.

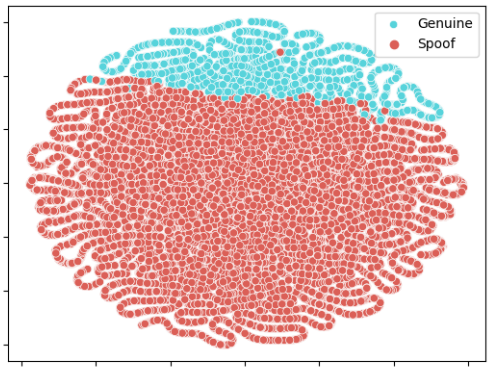

Figure 6.5: t-SNE plot for embeddings of ILPRgram+LPCgram feature on the evaluation set of ASVSpoof 2019 database

6.5 Source-Filter Time-Frequency Representation for RAD

In the previous section, we advocated using a deep residual network (ResNet-18) on our proposed CSFCC features. This network proved to model the proposed features much better than the GMM-based back-end, as evidenced by the RAD performances. The proposed features happen to be predominantly decorrelated due to the application of the DCT operations. However, even if correlated features are fed as input into a deep neural network, it can learn the appropriate feature transformations, and it is found that such features perform better than the decorrelated features. The authors of [44] argue that the power spectrum of speech contains relevant information about the artifacts introduced by the spoofing attacks. Motivated by that, they propose the use of time-frequency representation for spoof detection. In their study, different kinds of spectral representations are extracted using CQT, FFT, and DCT. The corresponding resultant features are called CQTgrams, FFTgrams, and DCTgrams. These features are fed into an LCNN. The CQTgram and LCNN-based RAD system performs the best. In the work done in [46], STFTgram and

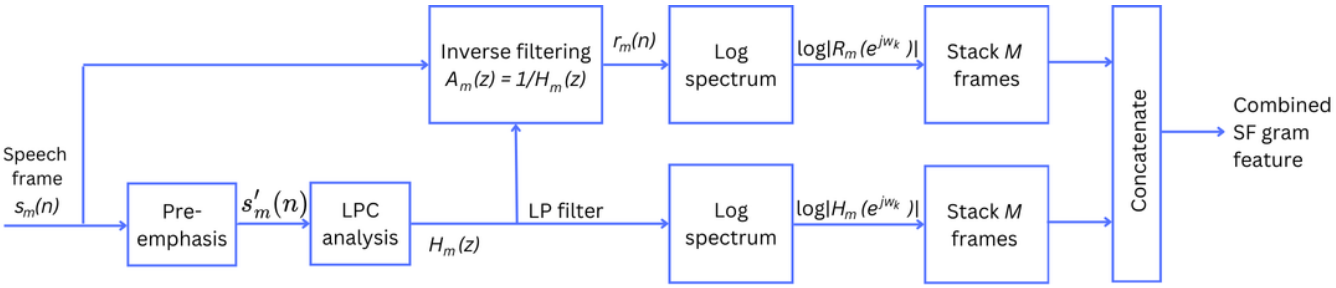

Figure 6.6: Block diagram showing the process of extracting the proposed combined source-filter gram(CSFgram) feature corresponding to a speech frame.

group delay grams are used along with a ResNet back-end, and the GDgrams-based RAD system outperformed all the other systems. Some gram-based features were also used in [45]. One of them was the CQTgrams that were extracted from the magnitude part of the spectrum. From the phase part of the spectrum, two more gram features were extracted, namely the modified group delay grams (MGDgram) and the CQT-based modified group delay grams (CQTMGDgram). The RAD systems developed using gram-based features with a modified ResNet called the ResNeWt were reported to perform the best in this work.

Motivated by the success of the gram-based features, we suggest a minor modification to the proposed CSFCC features. The 2D-DCT operations applied to both the source and filter parts are dropped, and the stacked frames from both parts are concatenated to obtain a new representation called the combined source filter gram (CSFgram) features. The process of extracting the CSFgram features from the speech signal is given in Figure 6.6. It can be observed from this figure that in the process of extracting the CSFgram features, we obtain two intermediate gram-based representations. One of them is from the source component of the signal and is referred to as the ILPRgram features. The second representation is generated from the filter component of the input signal and is termed the LPCCgram features.

6.5.1 Experimental Setup

We develop different RAD systems based on the gram-based features described above and the ResNet-18 classifier. The experimental setup of these systems is described next.

First, we describe the process of building a RAD system with the CSFgram feature. To extract the CSFgram feature, first, the input speech signal is short-term processed using a Hanning window

of size 20 ms and a shift of 16 ms. This choice of shift is made in order to reduce the number of frames so that the memory requirements to train the ResNet-18 model can be met. Next, an LP analysis of order 20 is performed on each windowed frame to obtain the LP coefficients. The LP coefficients are passed through two different filters, an inverse filter and an LP filter. First, the coefficients are passed through an inverse filter to obtain the source component and then a 512-point FFT is applied to it. Then the obtained spectrum is logarithmically compressed to get the log-spectrum of the source component. This is done for all the windowed frames and the log spectra are stacked together to obtain the ILPRgram feature. The size of the ILPRgram depends on the number of frames. If the number of frames for an input speech file is N, then the dimensionality of the feature is $N \times 512$.

As mentioned earlier, the LP coefficients are also passed through an LP filter and then 512-point FFT is applied to the output of the filter to calculate the spectrum. Then we apply log to the spectrum and stack the log spectra of all the frames to obtain the LPCCgram features. Similar to the ILPRgram, the dimensionality of the LPCCgram features is also $N \times 512$ where N is the number of frames. The ILPRgram and LPCCgram features are concatenated to obtain the final CSFgram features. The size of the CSFgram features is thus $2N \times 512$. The extracted CSFgram features are used to build a RAD system using a ResNet-18 back-end. The experimental setup of the ResNet-18 is the same as explained in Section 6.4.2.1. The intermediate features, ILPRgram and LPCCgram are also used to develop two more RAD systems with the same ResNet-18 back-end.

To benchmark the performance of the RAD systems developed using gram-based features we have extracted by tweaking the CSFCC features, we also build a RAD system with CQTgram features and the same ResNet-18 back-end. CQTgrams are extracted with a 32 ms frameshift, Hanning window, 11 octaves, and 48 bins per octave. Finally, another RAD system is built by combining the ILPRgram and CQTgram features as the front-end and ResNet-18 as the back-end.

6.5.2 Results and Discussion

The performances of the RAD systems using gram-based features and ResNet-18 back-end on the PA set of the ASVSpoof 2019 database are shown in Table 6.5. From this table, it can be seen that the CSFgram features perform much better than the CSFCC features on this database.

The CSFCC features provide an EER of 1.97% on the evaluation set of the database, whereas the CSFgram features result in an EER of 1.05%. This is a relative improvement of 97%.

The CQTgram-based RAD system gives an EER of 0.89% on the evaluation set of the ASVSpoof 2019 database. Finally, the feature-level combination of ILPRgram and CQTgram gives us the best performance obtained in this work in terms of both t-DCF and EER on the development and evaluation sets of this database. This confirms that the gram-based features are more suitable for the ResNet-18 back-end and that this system can be used as an effective RAD system.

Table 6.5: RAD performances of some gram-based features with Resnet-18 classifier on the ASVSpoof 2019 database

	Development		**Evaluation**	
Feature	**min-tDCF**	**EER (%)**	**min-tDCF**	**EER (%)**
LPCCgram	0.051	1.84	0.089	3.06
ILPRgram	0.035	1.22	0.069	2.55
CSFgram	0.012	0.44	0.030	1.05
CQTgram	0.008	0.28	0.024	0.89
ILPRgram + CQTgram	0.009	0.30	**0.022**	**0.72**

6.6 Comparison with Contemporary Features

In the previous section, it was demonstrated that the RAD system developed using the combined ILPRgram and CQTgram features along with the ResNet-18 classifier turned out to be the best system as evidenced by its performance on the ASVSpoof 2019 database. In this section, we present a comparison of that system to a few contemporary systems that have also shown superlative performance on the PA set of the ASVSpoof 2019 database. This comparison is shown in Table 6.6. The table illustrates the standing of the ILPRgram + CQTgram features in comparison to other features.

6.7 Conclusion

In this chapter, we propose augmenting source features with filter-based features. To this end, a new feature derived from the LP analysis called the 2D-LPCC feature is designed. The parameters of these features are tuned on the ASVSpoof 2017 v2.0 database. Then, these features are combined with the previously defined 2D-ILRCC features to obtain the proposed CSFCC features. The

Table 6.6: Survey of RAD system performances evaluated on ASVSpoof 2019 database.

System (feature, back-end)	Development		Evaluation	
	EER(%)	t-DCF	EER(%)	t-DCF
[31] (LFCC, GMM)	11.96	0.255	13.54	0.302
[31] (CQCC, GMM)	9.87	0.195	11.04	0.245
(2D-ILRCC, GMM)	9.15	0.199	11.35	0.272
[120](CQCC, ResNet)	4.30	0.103	4.43	0.107
[120] (Spectrogram, ResNet)	3.85	0.096	3.81	0.099
(2D-ILRCC, ResNet18)	1.62	0.044	2.33	0.065
[51] (CQT-MMPS, ResNet)	-	-	1.08	0.031
(ILPRgram + CQTgram, ResNet-18)	0.30	0.009	0.72	0.022

CSFCC features are evaluated on the PA set of the ASVSpoof 2019 database. It is found that the proposed CSFCC features outperform both the baseline CQCC and LFCC features. Further, the use of a deep residual network called ResNet-18 is advocated instead of the simple GMM as the back-end to develop RAD systems. Accordingly, a RAD system is developed with CSFCC features as the front-end and a ResNet-18 back-end. The back-end of the baseline RAD systems are also replaced with a ResNet-18. It is found that the CSFCC and ResNet-18-based RAD systems perform significantly better than the corresponding baseline systems in terms of both EER and t-DCF. Next, a modification to the CSFCC features is proposed where the spectrograms corresponding to the source and filter components are considered directly without the application of DCT for compression. The stacking of source and filter spectrograms along the temporal dimension is termed the CSFgram features. The CSFgram features with a ResNet-18 back-end are evaluated on the PA set of the ASVSpoof 2019 database. A contrast system is also developed with CQTgram features and a ResNet-18 back-end. It is found that the CSFgram features perform better than the CSFCC features. Moreover, the CSFgram features also outperform the CQTgram features. Finally, the ILPRgram and CQTgram features are combined at the feature level and a RAD system is developed with a ResNet-18 back-end. This system gives an EER of 0.72% on the ASVSpoof 2019 database and is the best system in this work.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

7

Summary and Future Directions

Contents

7.1 Summary of the Work

This thesis addresses the issue of detecting replay attacks in automatic speaker verification (ASV) systems. There are three major goals of the thesis. The first one is to establish that the source component of a speech signal serves as a possible alternative for characterizing replay attacks. To this end, different ways of processing the source signals are proposed resulting in the design of different kinds of replay attack detection (RAD) features. The second is to investigate the relative contribution of the voiced and unvoiced speech regions towards RAD. The final goal of the thesis is to examine the effect of jointly modeling the decomposed source and filter components of a speech signal for detecting replay attacks.

The thesis begins with an introduction to ASV and replay attacks. The task of an ASV system is to either accept or reject the identity claim of a speaker. Replay attacks are perpetrated on an ASV system by secretly recording the speech of a genuine user of the system and then playing it back to the system to gain unauthorized access. A detailed literature review highlighting the evolution of RAD is then provided. The literature review showed that the focus of the existing RAD works is on the direct characterization of replay attacks from the speech signal. However, no attempt has been made to decompose the speech signal into its constituent source and filter components and study the effects of replay attacks on these components separately. This thesis posits that both these components are affected by replay attacks and exploiting the spoofing information present in the two components can lead to better modeling of replay attacks. Accordingly, the thesis first delves into the exploration of source modeling for replay attacks to assess how much impact it has on RAD performance. Following this, the techniques developed to model the source information are enriched with the information contained in the filter component.

The various source and filter information modeling techniques used to develop the RAD systems are validated using two databases called the ASVSpoof 2017 v2.0 and the physical access (PA) set of the ASVSpoof 2019. A thorough description of these two databases is given in Chapter 2. Both these databases are divided into train, development, and evaluation subsets. With the ASVSpoof 2017 v2.0 database, we deal with the task of developing standalone RAD systems, and the performances of the systems are measured in terms of equal error rate (EER) and decision cost function (DCF). On the other hand, the task with ASVSpoof 2019 is to develop a RAD system in conjunction with

an ASV system. Hence, in addition to the EER, another metric called the tandem detection cost function (t-DCF) is utilized to measure the performance of the developed RAD system. This metric measures the cumulative performance of the RAD and ASV systems. These databases are used to develop two baseline RAD systems for this thesis. The first system uses the constant-Q cepstral coefficient (CQCC) features as the front-end and a Gaussian mixture model (GMM) based back-end. The second one utilizes linear frequency cepstral coefficients (LFCC) features and a GMM classifier to detect replay attacks.

It has been mentioned previously that this thesis aims to decompose the speech signal into source and filter components, and model replay attacks using these two components. The approach taken to achieve this is to first analyze the source signal for RAD and subsequently add the information present in the filter component for better characterization of replay attacks. Hence, Chapter 3 describes the preliminary work done to characterize replay attacks using the source signal. This chapter serves as the basis for the research conducted on source modeling for RAD. A review of the existing glottal source modeling techniques is presented initially. From this review, we choose to use the zero-frequency filtered (ZFF) signal and the linear prediction residual (LPR) signal as the two glottal source representations. A visual inspection of the ZFF and LPR signals for pairs of genuine and replayed speech. This study revealed that there are indeed differences in the nature of ZFF and LPR signals for genuine and replayed speech. Based on these observations, two handcrafted features are proposed. These are the epoch feature (EF) and the peak-to-side-lobe ratio mean and skewness (PSRMS). These features are used to develop two RAD systems with a GMM-based back-end. In addition, two cepstral features namely mel frequency cepstral coefficients (MFCC) and instantaneous frequency cepstral coefficients (IFCC) are also used to build different RAD systems for contrast purposes. All of these RAD systems are evaluated on the ASVSpoof 2017 v2.0 database. The performances of these systems reveal that the proposed source features yield decent RAD performance. This serves as empirical proof that source features also contain replay information. The EERs of the EF and PSRMS-based RAD systems on the evaluation set of the database are 32.51% and 28.9%, respectively. These performances are relatively poorer compared to the baseline CQCC and LFCC systems. This is attributed to the fact that the source features are naively handcrafted and are extracted only around the glottal closure instants (GCI). However,

the CQCC and LFCC features are extracted from the entire speech signal. Owing to this they are exposed to a much higher amount of information.

The handcrafted features capture information only around the GCIs and thus cannot extract features from other parts of the source signal. This is an inherent disadvantage of these features due to the way in which the source signal is processed for their extraction. Furthermore, the ZFF signal contains information only about the GCIs, and all other information is filtered out. The LPR signal, on the other hand, contains information along the entire source signal. However, it comprises multiple bipolar peaks which results in the detection of ambiguous GCIs. These drawbacks are addressed in Chapter 4 by introducing two new approaches. The first one is to use the integrated linear prediction residual (ILPR) signal to represent the source signal. The second approach is pitch-synchronous processing of the ILPR signal which captures the dynamics between two GCIs. This approach allows us to obtain replay information in the entire voiced regions of the signal instead of the neighbourhood of the GCIs. Combining these two approaches, a transform-based novel pitch-synchronous feature called the compressed ILPR (CILPR) is proposed for RAD. The CILPR feature captures the temporal details of the ILPR (source) signal between any two adjacent GCIs. It is observed that the temporal characteristics of the source signal differ in the replayed signal from those in the genuine signal due to the presence of replay artifacts. This discriminative information can be used for the effective detection of replay attacks. Hence, a RAD system is developed with the CILPR feature as the front-end and a GMM classifier as the back-end on the ASVSpoof 2017 v2.0 database. This system yields an EER of 20.66% on the evaluation set of the database. Compared to the PSRMS-based RAD system, we happen to get a relative improvement of 28.51%. Thus, highlighting the fact that the CILPR features happen to capture the replay attack distortions in a much better way than the handcrafted features. However, when compared to the baseline features, it results in significantly poorer RAD performance. This degradation in performance could be due to the use of only voiced regions of speech while extracting the CILPR features. An ablation study performed with a CQCC-GMM-based RAD system on the ASVSpoof 2017 v2.0 data showed that the unvoiced regions contained more RAD information than the voiced regions. Hence, in order to obtain comparable RAD performances, the source features should have been extracted from both the voiced and unvoiced regions.

The CILPR features, being extracted in a pitch-synchronous fashion, can be computed for the voiced regions only. The ablation study, on the other hand, revealed that the unvoiced regions also contain a significant amount of RAD information and can contribute to enhancing the performance of RAD systems. Hence, in Chapter 5, some modifications are proposed to overcome the shortcomings of the CILPR feature. The first modification is to adopt non-pitch-synchronous processing of the source (ILPR) signal through frame-based processing. This processing ensures that features can be extracted from the entire signal and also eliminates the need for pitch-marking. The second modification is to extract the features from a spectro-temporal domain instead of the temporal domain as is the case with the CILPR feature. This is done because it was found that there exists extended temporal continuity in the spectral structure of genuine speech compared to replayed speech. Thus, extracting features from the spectro-temporal domain can capture this discriminative information. Merging these two modifications, a new RAD feature called the two-dimensional integrated linear prediction residual cepstral coefficients (2D-ILRCC) is designed. This novel feature is used with a GMM back-end to develop two RAD systems on the ASVSpoof 2017 v2.0 database and the PA set of the ASVSpoof 2019 database. The results of the experiments show that the 2D-ILRCC system performs at par with the baseline systems on both databases. On the evaluation set of the ASVSpoof 2017 v2.0 database, the proposed 2D-ILRCC features yield an EER of 10.87% and on the PA set of the ASVSpoof 2019 database, it results in an EER of 11.35%.

The 2D-ILRCC-based RAD system provides competitive performance with respect to the baseline systems. However, it still fails to outperform the baseline systems. This is ascribed to the fact that the 2D-ILRCC features are extracted from only the source component of the speech signal, whereas the baseline CQCC features utilize the information present in both the source and filter components. Thus, in Chapter 6, the 2D-ILRCC features are enhanced by including the filter component information. LP analysis is performed to decompose the speech signal into the source and filter components. The resulting source component is the ILPR signal from which the 2D-ILRCC features are extracted. However, the LP coefficients obtained from the LP analysis have been ignored so far. Now, the LP coefficients are considered to represent the filter information, and using these coefficients a new feature called the two-dimensional linear prediction cepstral coefficients (2D-LPCC) is proposed. This feature is then concatenated with the 2D-ILRCC features to jointly

model the source-filter components. The concatenated features are named the combined source filter cepstral coefficients (CSFCC).

The dimensionality of the CSFCC features is first tuned by developing a RAD system with a GMM-based back-end on the ASVSpoof 2017 v2.0 database. Then the performance of the CSFCC-based RAD system is validated on the PA set of the ASVSpoof 2019 database. It is found that this RAD system provides an EER of 9.85% whereas the CQCC-based RAD system results in an EER of 11.66% on the evaluation set of the ASVSpoof 2019 database. This shows that the proposed CSFCC features provide much better RAD performance than the baseline features. Further, the CSFCC features are used to build another RAD system using a deep residual network (Resnet-18) back-end on the PA set of the ASVSpoof 2019 database. This system yields an EER of 1.97% and outperforms the CQCC-based RAD system on the same back-end that gives an EER of 4.35%. Finally, a modification is proposed to the CSFCC features where the compression done by the discrete cosine transform is dropped and the uncompressed spectro-temporal features are used directly for classification. These new features are called the combined source-filter gram (CSFgram) features and the RAD system developed with these features and the ResNet-18 network yields an EER of 1.05% on the PA set of the ASVSpoof 2019 database. Moreover, gram-based features are also computed for the decomposed source and filter components and are called ILPR gram and LPCC gram features. The ILPR gram features are combined with CQT gram features and a RAD system is developed on the same database. This system provides an EER of 0.72% and is the best-performing system of this thesis.

7.2 Contributions of the Thesis

The salient contributions of this thesis are listed below.

- Established that the source component of the speech signal contains significant replay attack information. Several source features were proposed which were used to develop different RAD systems and it was demonstrated that the source feature-based RAD systems were able to achieve performances comparable to the baseline systems. Source features served as a viable alternative for replay attack detection.

- Proposed the characterization of replay attacks via the decomposition of the speech signal into source and filter components. It was hypothesized that the replay attacks affect both the source and filter components and hence modeling the impact on both the components separately will lead to better characterization of replay attacks. The experiments conducted in the thesis supported this hypothesis.

- Studied the relative contribution of voiced and unvoiced regions of the speech for RAD. It was found that the unvoiced regions are affected more by replay attacks compared to the voiced regions. This is due to the higher distortion-to-speech ratio in the unvoiced regions. Thus, the unvoiced regions of speech contained more replay information than the voiced regions.

- Explored the joint modeling of the decomposed source and filter components of the speech signal for RAD. The previously proposed source features were augmented with a feature computed from the filter component. The combined source-filter feature-based RAD systems outperformed all previous systems. Thus, it established that both components are impacted by replay attacks, and combining the information present in these components leads to superior RAD performance.

7.3 Future Directions

This thesis primarily focuses on the decomposition of a speech signal into source and filter components for better modeling of replay attacks. To this end, a number of approaches have been proposed which have resulted in state-of-the-art performances. Despite this, there are a number of issues and challenges that can be addressed in the future. These possible future directions are:

- The first future direction that can be explored is the use of transformers as a back-end. In this work, the Resnet-18 has shown significant performance improvement due to the use of residual blocks. In recent times, transformers have been used in different areas for classification purposes. With the use of a multi-headed attention mechanism, transformers can lead to increased RAD performance.

- The features proposed in this work use the DFT for converting the time-domain signal to the frequency-domain. Although DFT has been effective for developing the features, it would be

interesting to see the impact of using a CQT in place of a DFT to compute the proposed features. CQT due to its nature of high spectral and temporal resolution may lead to better modeling of replay attacks.

- The next future direction that could be examined is the extension of the techniques proposed in this thesis to LA attacks and DeepFake attacks. In recent years, due to the rise of generative artificial intelligence (AI), DeepFake attacks have posed a severe challenge to ASV systems.
- Another possible direction to extend this work is to build a single system that can handle both replay attacks and impostor attacks on an ASV system. Such systems are called spoofing-aware ASV (SASV) systems. The techniques proposed here can be used directly to develop a SASV system and find out if they result in decent performances. The techniques can also be modified so that the SASV system becomes more efficient.

A

Detection of Glottal Activity Regions

Contents

A speech signal can broadly be divided into glottal and non-glottal regions depending upon the nature of excitation present in the region. When the vocal tract system is excited by vibrations of the vocal folds during the process of speech production the resulting speech segments are called glottal activity regions. These vibrations of the vocal folds generate a periodic or quasi-periodic excitation source signal. Glottal activity regions are characterized by energy, periodicity, and asymmetrical nature of the source signal [121], [122]. Thus glottal activity regions of speech can be identified by analyzing the features of the source signal. In this thesis, the detection of glottal activity regions is achieved using three different attributes of the voice source signal namely strength of excitation, normalized auto-correlation peak strength, and higher-order statistics. They are discussed in more detail in Section A.1, A.2 and A.3.

A.1 Strength of Excitation

During the production of voiced sounds, rapid movement of vocal folds occurs resulting in high energy during the closing phase of the glottis and giving higher strength to voiced speech around the epoch location. This strength around the epoch locations is called the strength of excitation (SoE) and can be computed by calculating the slope near the epoch locations of the ZFF signal. Mathematically, the strength of excitation $s_e[p]$ of the ZFF signal $z[n]$ at p^{th} epoch location can be derived as follows:

$$s_e[p] = \big|z[p+1] - z[p]\big| \tag{A.1}$$

Figure A.1(a) and (b) depicts a segment of a speech signal and its corresponding ZFF signal. The SoE evidence is shown in Figure A.1(c) and it can be clearly seen that the SoE values are high in the GA regions and low in the non-GA regions and hence can be used as a feature to detect GA regions.

A.2 Normalized Auto-Correlation Peak Strength

During the production of voiced sounds, the time-varying vocal-tract system is excited by quasi-periodic glottal pulses of air. The voice source signal is therefore quasi-periodic for voiced sounds. This nature of the source signal can be captured with the help of the normalized auto-correlation peak strength (NAPS) of the ZFF signal $z[n]$. The NAPS $n_p(\tau)$ for one frame of the ZFF signal

can be obtained as follows:

$$n_p(\tau) = \frac{\sum_{i=1}^{N} z[n] * z[n-\tau]}{\sum_{i=1}^{n} z^2[n]} \tag{A.2}$$

where τ is the lag and it represents the position of highest peak. Its value varies from 2.5 ms to 15 ms and represents periodicity. It is computed for each frame of 20 ms with a shift of 10 ms and interpolated. This evidence is depicted in Figure A.1(d). From the figure it is apparent that the value of NAPS is high is the GA regions when compared to that of the non-GA regions. It can also be used effectively in addition to SoE for detecting GA regions.

A.3 Higher Order Statistics

Significant excitation occurs at the glottal opening and closing instants during glottal activity. However, the strength during closing instants is more compared to that of opening instants due to differences in the airflow pressure around those instants. This results in asymmetric nature of the glottal pulses [123]. This property of the glottal signal can be captured by computing the skewness to kurtosis ratio (SKR) from the ZFF signal. Since this ratio may be influenced by high noise energy, appropriate power of skewness and kurtosis is considered to make the ratio a function of moments. The SKR is thus calculated according to the equation shown below:

$$\text{SKR} = \frac{\gamma^2}{\kappa^{1.5}} \tag{A.3}$$

where γ and κ denote the skewness and kurtosis of a frame of the ILPR signal respectively, and can be derived as:

$$\gamma = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(r[n]-\overline{r}\right)^3}{\left(\frac{1}{N}\sum_{n=1}^{N}\left(r[n]-\overline{r}\right)^2\right)^{\frac{3}{2}}} \tag{A.4}$$

$$\kappa = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(r[n]-\overline{r}\right)^4}{\left(\frac{1}{N}\sum_{n=1}^{N}\left(r[n]-\overline{r}\right)^2\right)^{2}} - 3 \tag{A.5}$$

where N is the number of samples in one frame of the ILPR signal $r[n]$ and $\overline{r}$ is the mean of $r[n]$.

The SKR is considered to be the higher-order statistics (HOS) feature in this work. It is calculated for a frame of 20 ms with a shift of 10 ms. This feature is shown in Figure A.1(e) and it can be observed that the HOS values are high in the GA regions and low in the non-GA regions.

A.4 Combination of the Three Evidences

The three evidences of SoE, NAPS, and HOS carry complementary information for different sound units [122]. Hence, combining the three types of evidence helps increase the accuracy of GA region classification. Combined evidence is calculated by first normalizing all three kinds of evidence to a maximum value and then averaging them. Figure A.1(f) shows the combined evidence and it can be observed that this evidence can correctly detect the GA regions. The final GA regions are depicted in red dotted lines.

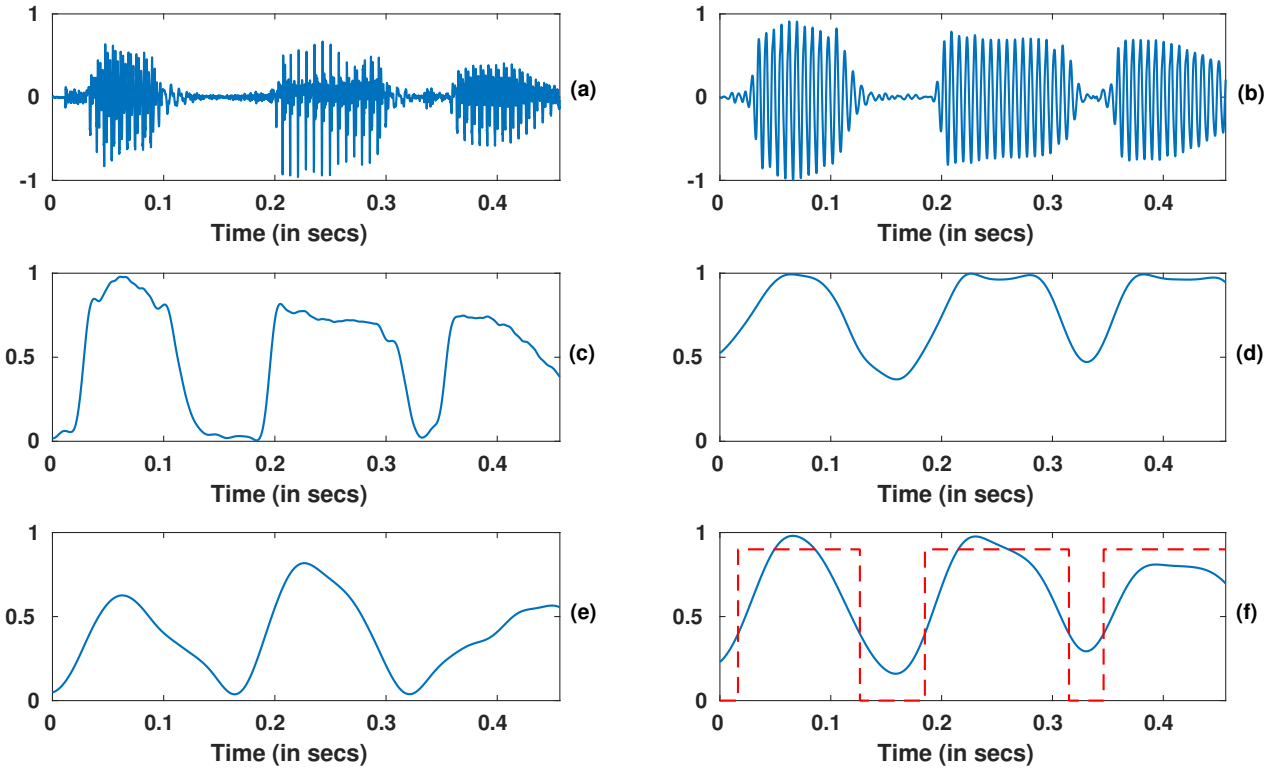

Figure A.1: Figure showing the process of extraction of glottal regions using three attributes of source information. (a) A segment of a speech signal, (b) the corresponding ZFF signal, (c) Glottal activity evidence using SoE (d) Glottal activity evidence using NAPS (e) Glottal activity evidence using HOS (f) Combined glottal evidence (blue line) and detected glottal regions marked with (red dotted line)

Bibliography

[1] A. Jain, A. Ross, and S. Pankanti, "Biometrics: A Tool for Information Security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

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

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[4] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000156

[5] "Types of Biometrics Voice Use Cases," https://www.biometricsinstitute.org/types-of-biometrics-voice-use-cases/.

[6] Y. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems, Springer*, 2005.

[7] P. L. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[8] C. B. Tan, M. H. A. Hijazi, N. Khamis, P. N. E. Nohuddin, Z. Zainol, F. Coenen, and A. B. Gani, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, pp. 32 725 – 32 762, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238768319

[9] 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.

[10] M. V. Mathews, J. E. Miller, and E. E. David, "Pitch synchronous analysis of voiced sounds," *Journal of the Acoustical Society of America*, vol. 33, pp. 179–186, 1961. [Online]. Available: https://api.semanticscholar.org/CorpusID:121213740

[11] J. Markel and A. G. Jr., "Linear prediction of speech," *Springer-Verlag, New York*, 1976.

[12] A. Oppenheim, G. Kopec, and J. Tribolet, "Signal analysis by homomorphic prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 327–332, 1976.

[13] I. Konvalinka and M. Matausek, "Simultaneous estimation of poles and zeros in speech analysis and itif-iterative inverse filtering algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 5, pp. 485–492, 1979.

[14] Y. Miyanaga, N. Miki, N. Nagai, and K. Hatori, "A speech analysis algorithm which eliminates the influence of pitch using the model reference adaptive system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 1, pp. 88–96, 1982.

[15] H. Morikawa and H. Fujisaki, "System identification of the speech production process based on a state-space representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 252–262, 1984.

[16] C.-Y. Chi and W.-T. Chen, "A novel adaptive maximum-likelihood deconvolution algorithm for estimating positive sparse spike trains and its application to speech analysis," in *IEEE International Workshop on Intelligent Signal Processing and Communication Systems*, 1992.

[17] C.-Y. Chi and J.-Y. Kung, "A new cumulant based inverse filtering algorithm for identification and deconvolution of nonminimum-phase systems," in *[1992] IEEE Sixth SP Workshop on Statistical Signal and Array Processing*, 1992, pp. 144–147.

[18] J. Tugnait, "Estimation of linear parametric models using inverse filter criteria and higher order statistics," *IEEE Transactions on Signal Processing*, vol. 41, no. 11, pp. 3196–3199, 1993.

[19] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification : A study of technical impostor techniques," in *EUROSPEECH 1999, 6th European Conference on Speech Communication and Technology*, 1999.

[20] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA*, 2010.

[21] Z. F. Wang, G. Wei, and Q. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *ICMLC*, 2011.

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

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

[24] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[25] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. G. S. Mello, R. P. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul, G. Saha, and M. Sahidullah, "Overview of btas 2016 speaker anti-spoofing competition," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–6.

[26] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[27] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech 2013*, 2013, pp. 925–929.

[28] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech 2015*, 2015, pp. 2037–2041.

[29] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Odyssey*, 2016.

[30] 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 *Proc. Interspeech 2017*, 2017.

[31] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. H. Kinnunen, and K. A. Lee, "ASVSpoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1008–1012.

[32] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[33] R. Font, J. M. Espin, and M. J. Cano, "Experimental analysis of features for replay attack detection–results on the asvspoof 2017 challenge," in *Interspeech*, 2017.

[34] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using dnn for channel discrimination," in *Interspeech*, 2017.

[35] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017," in *Interspeech*, 2017.

[36] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio Replay Attack Detection with Deep Learning Frameworks," in *Proc. Interspeech 2017*, 2017, pp. 82–86.

[37] T. Gunendradasan, B. Wickramasinghe, N. P. Le, E. Ambikairajah, and J. Epps, "Detection of replay-spoofing attacks using frequency modulation features," in *Proc. Interspeech 2018*, 2018, pp. 636–640. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1473

[38] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep siamese architecture based replay detection for secure voice biometric," in *Proc. Interspeech 2018*, 2018, pp. 671–675. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1819

[39] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *Proc. Interspeech 2018*, 2018, pp. 681–685. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2279

[40] A. T. Patil, R. Acharya, P. A. Sai, and H. A. Patil, "Energy Separation-Based Instantaneous Frequency Estimation for Cochlear Cepstral Feature for Replay Spoof Detection," in *Proc. Interspeech 2019*, 2019, pp. 2898–2902.

[41] T. Gunendradasan, E. Ambikairajah, J. Epps, and H. Li, "An Adaptive-Q Cochlear Model for Replay Spoofing Detection," in *Proc. Interspeech 2019*, 2019, pp. 2918–2922.

[42] M. Liu, L. Wang, J. Dang, S. Nakagawa, H. Guan, and X. Li, "Replay attack detection using magnitude and phase information with attention-based adaptive filters," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6201–6205.

[43] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on asvspoof 2019," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1018–1025.

[44] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," in *Interspeech*, 2019.

[45] X. Cheng, M. Xu, and T. F. Zheng, "Replay detection using cqt-based modified group delay feature and resnewt network in asvspoof 2019," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 540–545, 2019.

[46] W. Cai, Haiwei, D. Cai, and M. Li, "The dku replay detection system for the asvspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," *ArXiv*, vol. abs/1907.02663, 2019.

[47] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble Models for Spoofing Detection in Automatic Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 1018–1022.

[48] M. R. Kamble, H. Tak, and H. A. Patil, "Amplitude and frequency modulation-based features for detection of replay spoof speech," *Speech Communication*, vol. 125, pp. 114–127, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763932030279X

[49] P. Gupta, P. K. Chodingala, and H. A. Patil, "Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components," *Computer Speech Language*, vol. 77, p. 101423, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230822000559

[50] P. von Platen, F. Tao, and G. Tur, "Multi-Task Siamese Neural Network for Improving Replay Attack Detection," in *Proc. Interspeech 2020*, 2020, pp. 1076–1080.

[51] J. Yang, H. Wang, R. Das, and Y. Qian, "Modified magnitude-phase spectrum information for spoofing detection," *IEEE/ACM Transactions on Audi, Speech, and Language Processing*, February, 2021.

[52] Z. Lei, H. Yan, C. Liu, M. Ma, and Y. Yang, "Two-path gmm-resnet and gmm-senet for asv spoofing detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6377–6381.

[53] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 16–21.

[54] J. Cáceres, R. Font, T. Grau, and J. Molina, "The Biometric Vox System for the ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 68–74.

[55] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.

[56] Y. Lei, X. Huo, Y. Jiao, and Y. K. Li, "Deep Metric Learning for Replay Attack Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 42–46.

[57] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," in *Interspeech*, 2019.

[58] S. Yoon and H.-J. Yu, "Multiple-Point Input and Time-Inverted Speech Signal for The ASVspoof 2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 37–41.

[59] S.-H. Yoon and H.-J. Yu, "Multiple points input for convolutional neural networks in replay attack detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6444–6448.

[60] S.-H. Yoon, M.-S. Koh, and H.-J. Yu, "Phase Spectrum of Time-flipped Speech Signals for Robust Spoofing Detection," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 319–325.

[61] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[62] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[63] N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet V2: practical guidelines for efficient CNN architecture design," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11218. Springer, 2018, pp. 122–138.

[64] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 2820–2828. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Tan_MnasNet_Platform-Aware_Neural_Architecture_Search_for_Mobile_CVPR_2019_paper.html

[65] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 296–303.

[66] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech Language*, vol. 64, p. 101114, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300474

[67] K. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, M. J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Interspeech, Annual Conf. of the Int. Speech Comm. Assoc*, 2015.

[68] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:213060286

[69] T. H. Kinnunen, K. A. LEE, H. Delgado, N. W. D. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, vol. abs/1804.09618, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:13745361

[70] D. Gabor, "Theory of communication," *Journal of the Institution of Electrical Engineers*, vol. 93, pp. 429–457, 1946.

[71] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. BRILL, 2003.

[72] J. Youngberg and S. Boll, "Constant-q signal analysis and synthesis," in *ICASSP '78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, April 1978, pp. 375–378.

[73] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, pp. 425–434, 1991.

[74] M. Todisco, H. Delgado, and N. Evans, "Constant-Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, 2017.

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

[76] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[77] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.*, vol. 10, pp. 19–41, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:9760419

[78] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[79] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[80] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[81] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech Language*, vol. 60, p. 101027, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819302712

[82] 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, phase-Aware Signal Processing in Speech Communication.

[83] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech Language*, vol. 28, no. 5, pp. 1117–1138, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230814000229

[84] G. Fant, *Acoustic Theory of Speech Production*. Berlin, Boston: De Gruyter Mouton, 1971. [Online]. Available: https://doi.org/10.1515/9783110873429

[85] G. Fant, J. Liljencrants, Q.-g. Lin *et al.*, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[86] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971. [Online]. Available: https://doi.org/10.1121/1.1912389

[87] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, 1986, pp. 1605–1608.

[88] R. Veldhuis, "A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 566–571, 01 1998. [Online]. Available: https://doi.org/10.1121/1.421103

[89] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[90] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, 1992, eurospeech '91. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939290005R

[91] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 492–501, 2006.

[92] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, 2005.

[93] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech Language*, vol. 26, no. 1, pp. 20–34, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230811000210

[94] ——, "Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639311000227

[95] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 369–377, 1985.

[96] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998.

[97] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov 2008.

[98] N. Adiga and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in *INTERSPEECH*, 2013.

[99] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

[100] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 751 – 761, 2005.

[101] B. Sharma and S. R. M. Prasanna, "Sonority measurement using system, source, and suprasegmental information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 505 – 518, 2017.

[102] S. L. Marple, "Computing the discrete-time analytic signal via fft," in *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136)*, vol. 2, Nov 1997, pp. 1322–1325.

[103] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Chapter 3 - features for content-based audio retrieval," in *Advances in Computers: Improving the Web*, ser. Advances in Computers. Elsevier, 2010, vol. 78, pp. 71–150. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065245810780037

[104] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.

[105] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *JASA EL*, 2015.

[106] R. K. Das and S. R. M. Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016. [Online]. Available: https://doi.org/10.1121/1.4954653

[107] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4733–4736.

[108] V. Karjigi and P. Rao, "Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling," *Speech Communication*, vol. 54, no. 10, pp. 1104–1120, 2012.

[109] S. Strömbergsson, G. Salvi, and D. House, "Acoustic and perceptual evaluation of category goodness of/t/and/k/in typical and misarticulated children's speech," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3422–3435, 2015.

[110] S. Kalita, S. Mahadeva Prasanna, and S. Dandapat, "Intelligibility assessment of cleft lip and palate speech using gaussian posteriograms based on joint spectro-temporal features," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. 2413–2423, 2018.

[111] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing Detection on the ASVspoof2015 Challenge Corpus Employing Deep Neural Networks," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2016)*, 2016, pp. 270–276.

[112] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, pp. 43–52, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639316301091

[113] L. Huang and C.-M. Pun, "Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced densenet-bilstm network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1813–1825, 2020.

[114] H. Dinkel, Y. Qian, and K. Yu, "Investigating raw wave deep neural networks for end-to-end speaker spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2002–2014, 2018.

[115] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[116] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, "The SJTU Robust Anti-Spoofing System for the ASVspoof 2019 Challenge," in *Proc. Interspeech 2019*, 2019, pp. 1038–1042.

[117] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1003–1010.

[118] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[119] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[120] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Proc. Interspeech 2019*, 2019, pp. 1078–1082.

[121] N. Adiga and S. R. M. Prasanna, "Detection of glottal activity using different attributes of source information," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2107–2111, 2015.

[122] N. Adiga, B. K. Khonglah, and S. Mahadeva Prasanna, "Improved voicing decision using glottal activity features for statistical parametric speech synthesis," *Digital Signal Processing*, vol. 71, pp. 131–143, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200417302105

[123] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Publications Related to the Thesis

Journal Publications

1. **Sarfaraz Jelil**, Rohit Sinha, and S. R. Mahadeva Prasanna "Replay Attack Detection via the Decomposition of the Speech Signal into Source and Filter Components," To be submitted to *Computer Speech and Language*.

1. **Sarfaraz Jelil**, Rohit Sinha, and S. R. Mahadeva Prasanna "Spectro-Temporally Compressed Source Features for Replay Attack Detection," *IEEE Signal Processing Letters*, vol. 31, pp. 721-725, 2024.

Conference Publications

1. **Sarfaraz Jelil**, Sishir Kalita, S. R. Mahadeva Prasanna, and Rohit Sinha, "Exploration of Compressed ILPR Features for Replay Attack Detection," in *Proceedings of Interspeech*, 2018.

2. **Sarfaraz Jelil**, Rohan Kumar Das, S. R. Mahadeva Prasanna, and Rohit Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features," in *Proceedings of Interspeech*, 2017.

List of Publications Not Related to the Thesis

Journal Publications

1. Rohan Kumar Das, **Sarfaraz Jelil**, and S. R. Mahadeva Prasanna, "Multi-style Speaker Recognition Database in Practical Conditions," *International Journal of Speech Technology*, vol. 21, pp. 409-419, 2018.

2. Rohan Kumar Das, **Sarfaraz Jelil**, and S. R. Mahadeva Prasanna, "Development of Multi-level Speech-based Person Authentication System," *Journal of Signal Processing Systems*, vol. 88, pp. 259–271, 2017.

Conference Publications

1. **Sarfaraz Jelil**, Rohan Kumar Das, S. R. Mahadeva Prasanna, and Rohit Sinha, "Role of Voice Activity Detection Methods for the Speakers in the Wild Challenge," in *Proceedings of the National Conference on Communications*, 2017.

2. Rohan Kumar Das, **Sarfaraz Jelil**, and S. R. Mahadeva Prasanna, "Significance of Constraining Text in Limited Data Text-Independent Speaker Verification," in *Proceedings of Signal Processing and Communications*, 2016.

3. Rohan Kumar Das, **Sarfaraz Jelil**, and S. R. Mahadeva Prasanna, "Exploring Session Variability and Template Aging in Speaker Verification for Fixed Phrase Short Utterances," in *Proceedings of Interspeech*, 2016.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati