Design of Countermeasures for Spoofed Speech Detection System

by

TANVINA B. PATEL

201221003

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

DOCTOR OF PHILOSOPHY

in

INFORMATION AND COMMUNICATION TECHNOLOGY

to

DHIRUBHAI AMBANI

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

April, 2017

Declaration

I hereby declare that

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

ii. Due acknowledgement has been made in the text to all the reference material used.

_____

Tanvina Bhupendrabhai Patel

Certificate

This is to certify that the thesis work entitled, *"Design of Countermeasures for Spoofed Speech Detection System"*, has been carried out by *Tanvina B. Patel (201221003)* for the degree of Doctor of Philosophy in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) under my supervision.

_____

Prof. (Dr.) Hemant Arjun Patil

Thesis Supervisor

To

My Family

Acknowledgments

"Words are never sufficient to express the amount of gratitude that a person deserves; but still… words are the only way we can express our feelings."

"Good supervisors can take you to incredible heights. They help you learn to soar, providing the wind underneath you, and providing the support for when you fall". Hence, first and foremost, I present my greatest gratefulness to Prof. Hemant A. Patil (DA-IICT, Gandhinagar) whose guidance laid the strong foundation for my thesis. He has been extremely supportive in listening to my queries and providing ample time for discussions. I thank him for perilously going through all my research papers and giving suggestions for improvements. It was his support that made it possible to submit my work to excellent conferences on time. With his motivation and moral support till submission, I could see the completion of this thesis.

I thank my Research Progress Seminar (RPS) and Synopsis committee members: Prof. M. V. Joshi, Prof. S. K. Mitra, Prof. L. Pillutla and Prof. Y. Yasavda for their insightful comments and encouragement, and also for their questions which helped me to widen my research from various perspectives. I thank the authorities of DA-IICT, Gandhinagar for their support to carry out the research work. I acknowledge the suggestions given by the thesis examiners, Prof. Nicholas Evans, Eurecom, France and Prof. Sriram Ganapathy, IISc, Bangalore for the various suggestions in the thesis. I also thank the anonymous reviewers for their feedbacks that helped to strengthen our manuscripts for acceptance and in turn helped in the making of this thesis. I take this opportunity to thank the organizers and the entire team of the ASV spoof 2015 challenge for making the ASV spoof challenge database available without which such intensive work in the thesis would not be inevitable at all.

I acknowledge the support and funding from Department of Electronics and Information Technology (DeitY), Govt. of India, New Delhi, for the consortium project, "Development of Text-to-Speech Synthesis Systems in Indian Languages-Phase II". I appreciate the efforts of consortium leader Prof. Hema Murthy (IIT, Madras). It was through DeitY that the speech lab at DA-IICT was developed and the resources were available. Also, I thank the support from International Speech

Communication Association (ISCA), Microsoft Research (MSR) and IEEE Signal Processing Society (SPS) for providing partial travel grants to present research papers at INTERSPEECH 2013, INTERSPEECH 2014 and ISCSLP 2014 and ICASSP 2016, respectively. Attending and presenting at the conferences have helped me gain international exposure and I had an altogether different perspective of my research. The Prosody and TTS Workshops, Summer and Winter Schools at DA-IICT and elsewhere have helped me to stay connected with the speech group in India.

My special gratitude and sincere thanks to all the members of the speech research lab at DA-IICT. Starting with Maulik bhai, thank you for the intensive practical knowledge you have been sharing with all the members. Anshu Di, I appreciate the time you gave even with the personal commitments. Yours was the first thesis which prepared me mentally for mine. Next are the two respected MoU buddies Nirmesh and Hardik with whom I worked jointly for *four* years in the TTS project. The technical discussion we have had has helped me tremendously in enriching my research work. Definitely, do not expect me to forget the times you quarreled with me! I thank Swati, Zaki, Janki and Vibha for being so committed to the TTS project when they were a part of it. I appreciate the support of all TTS consortium members for the technical guidance and their efforts towards the consortium paper. I also acknowledge the contributions of other past members and my co-authors at the lab, Pramod, Avni, Deep, Meet, Himanshu, for working together. The new members of the lab, Ashish, Ankur, Apeksha, Dharmesh, Madhu and Rishabh, thanks for biding me the farewell.

I again acknowledge the speech lab at DA-IICT, this time as a second home. With Shubham being a part of the family, followed by the M.Tech. batch of Avni, Pramod, Purvi, and Ankit. It was because of you all I realized that there is a fun side of Ph.D. life too. I do not forget the gang of girls, Ami, Pratika, Priyanka, Neha, Swati and Ilam. The final stages of Ph.D. are also the tough ones, especially when you cannot express and yet want someone to understand. Thank you, Avni and Rahul for being there during my synopsis and thesis submission. I know, you talked or constantly blabbered at times, just to let the stress off me. Pramod, although now on a foreign land, you would have done the same. Thank you all for the TDi tag. I always cherish the moments spent with all of you. At times it was necessary to be out of speech and thanks to my non-speech people at DA-IICT, Trupti Di (my roomie), Dixita and

Kamal who have been excellent colleagues and friends. To my mentors and friends off campus who were eagerly waiting for my thesis, I thank for the best wishes that have always been with me.

Lastly, with the least number of lines dedicated in this acknowledgement Section, I would like to thank my parents who raised me with love and supported me to pursue my dream in spite of several hardships. I may not have understood you at times, but you always did. And I am proud of you. My brother Devious who always behaves as his name is, but is silently wishing that I achieve what I deserve. My Ph.D. was possible because of the constant support of my fiancé Khushal who has left no stone unturned to make me reach where I am today. It was because of him that I could withstand all troubles that came my way in the final days of my Ph.D. I see this thesis coming up just because of my family, their trust in me, their patience and because of all the blessing they showered on me.

Table of Contents

Abstract

Automatic Speaker Verification (ASV) systems are vulnerable to speech synthesis and voice conversion techniques due to spoofing attacks. Recently, to encourage the development of anti-spoofing measures or countermeasures for Spoofed Speech Detection (SSD) task, a standardized dataset was provided at the 'ASV spoof 2015 challenge' held at INTERSPEECH 2015. In the present work, using a traditional Gaussian Mixture Model (GMM)-based classification system, novel countermeasures are proposed considering three vital aspects of speech production mechanism, i.e., excitation source, vocal tract system (i.e., filter) and Source-Filter (S-F) interaction.

Considering our relatively best performance at the ASV spoof challenge, we first discuss *system-based* features that include proposed Cochlear Filter Cepstral Coefficients and Instantaneous Frequency (CFCCIF) features. These use the envelope and average IF of each subband along with the transient information. The transient variations estimated by the symmetric difference (CFCCIFS) gave better discrimination. Within the framework of system-based features, the Subband Autoencoder (SBAE) feature set that embeds subband processing in the Autoencoder architecture is used. For *source-based* features, knowing that an actual vocal fold movement is absent in machine-generated speech, fundamental frequency (F_0) contour and Strength of Excitation (*SoE*) are used as features. Next, as spoofed speech is easily predicted if generated by a simplified model or difficult to predict due to artifacts, we propose the use of prediction-based methods. This includes the Linear Prediction (LP), Long-Term Prediction (LTP) and Non-Linear Prediction (NLP) techniques. Lastly, the Fujisaki Model is used to analyze the prosodic differences in terms of accent and phrase between natural and spoofed speech. In addition to independently using source or system features, the time-varying dependencies or the S-F interaction features are considered. This includes exploring features based on the residual information of the glottal excitation source and its fitted Liljencrants-Fant (LF) model, both in time-domain and frequency-domain for the SSD task.

Overall, the system-based features worked well for unknown attacks, especially the vocoder-independent spoof. On the other hand, source-based features when used

with the system-based features performed well for vocoder-based spoofs. Hence, the score-level fusion of system and source-based features significantly reduced the % Equal Error Rate (EER). The S-F interaction features were found to perform well for vocoder-based spoofs and were robust to signal degradation conditions as well. The performance of all features has been evaluated for the known, unknown, same and different type of attacks. Amongst the countermeasures existing in the literature, majority of them specifically model the artifacts introduced while synthesis or conversion. In the present work, we propose features that are characteristics of natural speech and difficult to incorporate into machine-generated speech. Finally, the features are also evaluated on the Blizzard Challenge 2012 and Blizzard Challenge 2014 database to study the channel mismatch effects in the SSD task. Although significant research is done to detect spoofed speech, the problem is yet to be completely solved for unit-selection synthesis, signal degradation and channel mismatch conditions.

List of Principal Symbols and Acronyms

Symbols

$s(t)$	Speech signal
$s(n)$	Discrete-time speech signal
$S(\omega)$	Spectrum of the speech signal $s(n)$
$h(t)$	Impulse response of the vocal tract filter
Fs	Sampling frequency
e_{LP}	Linear prediction residual
e_{LTP}	Long-term prediction residual
e_{NLP}	Non-linear prediction residual
$g(t)$	Glottal flow waveform
$\dot{g}(t)$	Glottal flow derivative waveform
$g_c(t)$	Coarse structure of $\dot{g}(t)$
F_0	Fundamental frequency
τ_{gd}	Group delay function
fc	Center frequency
μ	Mean
sd	Standard deviation
a_f	Fusion factor of score-level fusion
λ_{nat}	Gaussian mixture model for natural speech
λ_{spoof}	Gaussian mixture model for spoofed speech
$\psi_{a,b}(t)$	Cochlear filter
$W(a,b)$	Auditory transform
σ	Longitudinal stress
ρ	Density
Δ	Delta
$\Delta\Delta$	Delta-delta
Ds	Static features
$D1$	Static+delta
$D2$	Static+delta+delta-delta

Acronyms

AE	Autoencoder
ASpR	Automatic Speaker Recognition
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
AT	Auditory Transform
BM	Basilar Membrane
CART	Classification And Regression Trees
CF	Characteristic Frequency
CFCC	Cochlear Filter Cepstral Coefficients
CFCCIF	Cochlear Filter Cepstral Coefficients Plus Instantaneous Frequency
CFCCIFS	CFCCIF With Symmetric Difference
D	Dimension
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DET	Detection Error Tradeoff
DMOS	Degradation Mean Opinion Score
DNN	Deep Neural Network
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
FAR	False Acceptance Rate
FRR	False Rejection Rate
FS	Frame Selection
GD	Group Delay
GMM	Gaussian Mixture Model
GMM-UBM	Gaussian Mixture Model-Universal Background Model
GV	Global Variance
HAS	Human Auditory System
HFR	High Frequency Region
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
HSMM	Hidden Semi-Markov Model
HTS	HMM-based Speech Synthesis System
IAIF	Iterative Adaptive Inverse Filtering
IEC	International Electrotechnical Commission
IF	Instantaneous Frequency
IFD	Intraframe Differences
IHC	Inner Hair Cells
IPA	International Phonetic Alphabet
ISO	International Organization for Standardization

JFA	Joint Factor Analysis
KPLS	Kernel Partial Least Square
LBP	Local Binary Patterns
LF	Liljencrants-Fant
LFCC	Linear Frequency Cepstral Coefficients
LFR	Low Frequency Region
LLR	Log-Likelihood Ratio
LMS	Log-Magnitude Spectrum
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LSD	Log Spectral Distortion
LSP	Line Spectral Pair
LTI	Linear Time-Invariant
LTP	Long-Term Prediction
LTS	Letter-to-Sound
LTS	Letter-to-Sound
MARY TTS	Modular Architecture for Research on speech sYnthesis (MARY) Text-To-Speech
MBE	Multi-Band Mixed Excitation
MCC	Mel Cepstral Coefficients
MCD	Mel Cepstral Distortion
MFCC	Mel Frequency Cepstral Coefficients
MFPC	Mel Frequency Principal Coefficients
MGDF	Modified Group Delay Function
ML	Maximum Likelihood
MLSA	Mel Log Spectrum Approximation
MM	Magnitude Modulation
MOS	Mean Opinion Score
MPS	Mean Pitch Stability
NIST	National Institute of Standards and Technology
NLP	Non-Linear Prediction
nP	No Pre-emphasis
P	Pre-emphasis
PLDA	Probabilistic Linear Discriminant Analysis
PM	Phase Modulation
PMVDR	Perceptual Minimum Variance Distortionless Response
PSOLA	Pitch-Synchronous Overlap and Add
PSP	Pitch Synchronous Phase
RBM	Restricted Boltzmann Machine
RLMS	Residual Log Magnitude Spectrum
RMSE	Root Mean Square Error
RPS	Relative Phase Shift
S	Spoofing

SAS	Spoofing and Anti-Spoofing
SBAE	Subband Autoencoder
SF	Shrillness Factor
SID	Speaker Identification
SNR	Signal-to-Noise Ratio
SoE	Strength of Excitation
SPSS	Statistical Parametric Speech Synthesis
SS	Synthetic Speech
SSD	Spoofed Speech Detection
STFT	Short-Time Fourier Transform
STM	Spectral Transition Measure
STRAIGHT	Speech Transformation And Representation using Adaptive Interpolation Weighted Spectrum
SUS	Semantically Unpredictable Sentences
SV	Speaker Verification
SVM	Support Vector Machine
TEO	Teager Energy Operator
TTS	Text-to-Speech
TVC	Tensor-based Arbitrary Voice Conversion
USS	Unit Selection Synthesis
VCS	Voice Converted Speech
WSF	Window Scale Factor
WSJ	Wall Street Journal
ZF	Zero Frequency

List of Figures

List of Tables

List of Tables

Chapter 1.
Introduction

1.1 Introduction

Speech is a natural and powerful form of communication between individuals. It is very natural to produce and it is an inherent attribute or identity of an individual [**1**]. Speech is a signal in which the speech samples are changing dynamically over a period of time. It can be used as an adequate biometric modality, especially due to its remote access and convenience. Traditionally, from the deployment perspective, to facilitate the machines for authentication, biometrics such as fingerprints, face, iris, handwriting (such as signature), etc. are generally used for identification and verification tasks [**2**]. However, humans can identify and discriminate amongst speakers using the acoustic cues from the speech signal. In reality, humans by nature use face and voice biometrics jointly to identify an individual. The Automatic Speaker Verification (ASV) technology uses speaker-specific information from the speech signal for authentication, wherein the ASV system either accepts or rejects a claimed speaker's identity [**3**]. Increased use of ASV systems as a biometric has demanded or questioned its reliability under *spoofing* scenarios. That is, the ASV systems must be secure or robust against an adversary, generally referred to as an impostor who might try to deceive the voice biometric system by claiming as another user. The claim by the impostor can be done either by impersonation (or mimicry), replay or manipulating and generating speech signal artificially (such as synthetic speech or voiced converted speech).

Spoofing attacks are also known as *presentation* attacks as per the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) standardization [**4**]. According to the ISO/IEC, presentation attacks refer to "Presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system. Presentation attacks can be implemented through a number of methods (for example, artifact, mutilations, replay, etc.). The biometric systems may not be able to differentiate between

biometric presentation attacks with the goal of interfering with the systems operation and non-conformant presentations" [**4**]. The measures or systems that can detect the spoofing attacks (i.e., presentation attacks) from normal presentations are known as presentation attack measures or anti-spoofing measures or *countermeasures*. Presentation attack detection is also similar to liveness detection, i.e., "measurement and analysis of anatomical characteristics or involuntary or voluntary reactions, in order to determine if a biometric sample is being captured from a living subject present at the point of capture". Liveness detection methods are a subset of presentation attack detection methods [**4**].

This thesis proposes suitable measures (i.e., countermeasures or anti-spoofing measures) to detect whether the claimed speech is genuine speech or from an impostor representing spoofed speech. The Spoofed Speech Detection (SSD) task is certainly important due to the fact that the interest in biometric applications has grown significantly and hence, the biometric system should be able to detect malicious attacks. The ASV systems can aid in access control to physical facilities, computer-related web services or telephone resetting of passwords, etc. The assured performance of ASV system is needed for use in telephone banking transactions, electronic banking, and e-commerce. Users generally remember keywords known as passwords to access a particular utility. Using the same password is risky and also, it can be forgotten or stolen. This brings into the need of text-independent biometric systems to address this problem. Evidently, this can happen only if the biometric system is accurate with very low Equal Error Rates (EER) and also *reliable* to impostor attacks as well. An EER is an operating point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) is equal (which will be discussed in detail in Chapter 3). Thus, in this thesis, we propose the design of features or countermeasures for a spoof detector system that can identify natural *vs.* spoofed speech. The features should enhance differences between natural and spoofed speeches and should be independent of differences within the natural speech due to the size and shape of the vocal tract, larynx size, etc. In this thesis work, distinctive features are proposed to design a spoof detector system that can be used in future along with the ASV system for secure speaker authentication task.

1.2 Architecture of the ASV Systems

The Automatic Speaker Recognition (ASpR) can be operated in two modes, namely, Speaker Identification (SID) and Speaker Verification (SV). In the identification task, the system tries to identify who the test speaker is from the available set of speakers. On the other hand, in the verification task, the system verifies if the claimed identity is a true (genuine speaker) or from an impostor. Furthermore, ASpR can be further classified into text-dependent, vocabulary-dependent (or text-prompted) or text-independent systems [**5**]. In text-dependent case, the speaker utters only a specific text to be identified or verified. Such a system has an advantage of higher accuracy as the dependency due to varying text is reduced. Vocabulary-dependent systems are at a slightly higher-level where the speech is limited to a specific-domain such as digits, alphabets, etc. and the test phrase can be selected as a combination of the limited vocabulary. This kind of text-prompted SV was one of the first approaches to alleviate spoofing attacks [**6**]. The other extreme end includes a text-independent system that includes no bound on the text used for the ASpR system. As compared to the text-dependent case, the text-independent is more secure as text-dependent systems can be fooled easily if the test utterance is known. The text-independent systems are also flexible in terms of changing the test phrase for preventive measures against impostor attacks.

For the machines, the task of identification and verification can be viewed as a pattern recognition problem. In SID, the task is to classify patterns (in the form of feature vectors of the speech signal) in the test to one of the previously known patterns. For SV, the sample pattern of an unknown pattern together with the claimed identity is given. The task is to determine whether the sample pattern is sufficiently similar to the reference pattern associated with the claimed identity in order to accept or reject the claim. The application of SID is limited in the sense that the decision of identification can be made only if the test speaker is enrolled in the system, i.e., the specific speaker has been used to train the SID system. On the other hand, verification requires validating a speaker among the large group of speakers that may be unknown to the system. In speaker forensics, it is common to first perform an identification process to create a list of "best matches" and then perform a series of verification processes to determine a conclusive match. In addition, the

term *voice comparison* is much appropriate rather than voice recognition in speaker forensics applications [**7**].

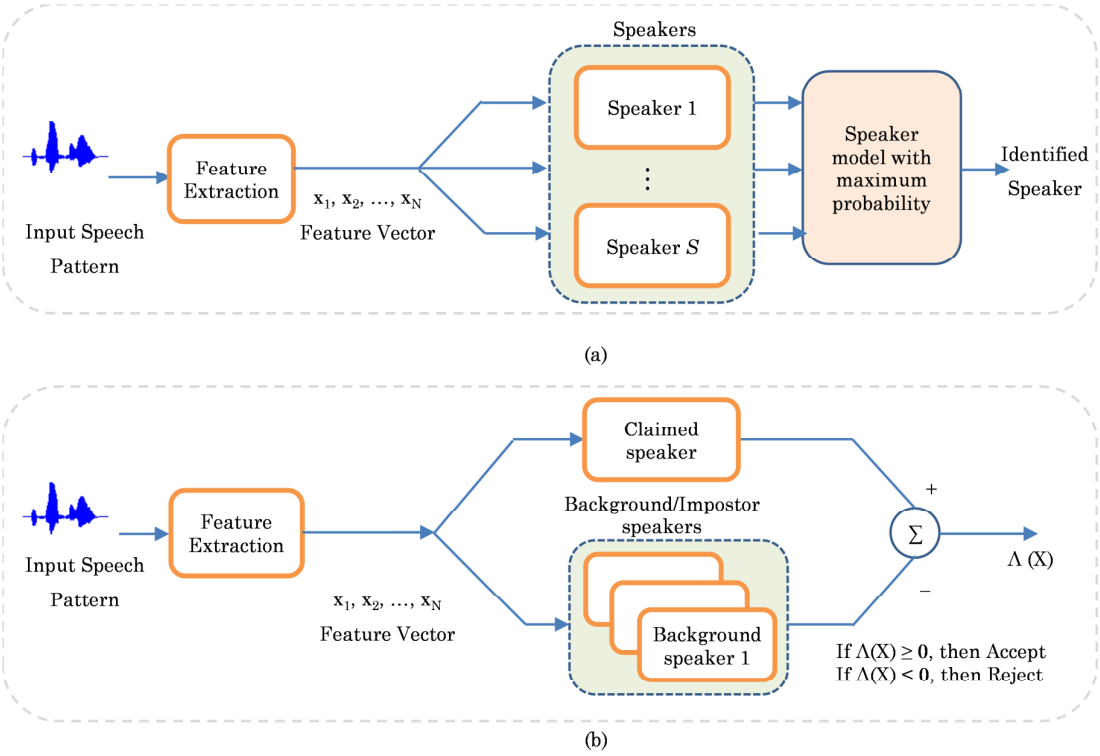


Figure 1.1: The speaker recognition systems (a) Speaker Identification (SID) and (b) Speaker Verification (SV). After [5].

As shown in Figure 1.1, features extracted from the input speech signal are given to the SID and SV systems [**5**]. These features can be either high-level, medium-level or low-level features. The high-level features include idiosyncrasies, diction, etc. which are peculiar to an individual. These features capture information such as the socio-economic background of the speaker. The medium-level features include prosodic, rhythm and intonation features. On the other hand, low-level cues are related to acoustic measurements that are directly related to the speaker's physiological characteristics (in particular, the size and shape of the vocal tract). While humans use all cues to verify a speaker, the recognition systems work well for low-level acoustic features (due to the practical difficulty of collecting hours of speech data from every speaker to extract meaningful high-level and medium-level features). The low-level features capture the physiological attributes of the speaker. As the vocal tract shape and size consists of most of the information of the speaker, spectral features depicting the resonances in the vocal tract are used. This spectral

representation includes passing speech through a set of subband filters of frequency ranges similar to that of the subband processing in the human ear (e.g., Mel-scale filterbank). These features are very well known in speech processing literature and are referred to as Mel Frequency Cepstral Coefficients (MFCC) [**8**]. In the identification systems, to create a statistical model, each speaker is considered as a random source generating the observed feature vector. Given the feature vectors, a statistical model such as Gaussian Mixture Model (GMM) is built for each speaker in the identification system. As shown in Figure 1.1 (a), in identification task, for a given test utterance, the speaker model with the maximum likelihood is considered as the speaker model of the test identity. On the other hand, for verification task as shown in Figure 1.1 (b), if the likelihood ratio of the test speaker or the claimant is greater than a certain threshold, the identity claim is accepted otherwise it is rejected. Thus, the entire speaker recognition technology can be summarized of voice recording (i.e., data collection and corpus design [**9**]), feature extraction, pattern matching (model training and likelihood estimation) and decision making. Research in ASpR has increased many folds with the advancement in various feature extraction techniques, use of Gaussian Mixture Model-Universal Background Model (GMM-UBM) modeling approaches, GMM supervectors, Support Vector Machine (SVM), *i*-vector, GMM with Joint Factor Analysis (JFA) [**10**] and Probabilistic Linear Discriminant Analysis (PLDA) [**11**]. The research challenges includes microphone errors (time-varying microphone placement), speaker variability (e.g., the vocal tract resonances may change with age and hence, alter the models, health conditions such as cold may change the voice quality of the speaker), intersession and microphone variability, robustness in presence of channel noise and most importantly robustness to spoofing attacks by the impostors.

1.3 Spoofing in ASV Systems

In 1976, an excellent review work by Rosenberg on ASV discussed the inclination towards verification systems than identification systems [**3**]. It states that SV is a more tractable problem where only a single comparison to a reference pattern is required which is faster and less complex. Hence, it makes the SV systems useful for several practical or commercial applications. The same review paper then discusses an important issue of impersonation by humans or the mimic *resistance* capacity of

verification systems (Section III-E, pp. 480 [**3**]). The mimic resistance is the property of the verification systems to resist determined mimics. Mimics can be based on physiological characteristics such as identical twins, or it can be based on behavior or learned characteristics such as professional mimics. Hence, the issue of dealing with mimics is essential as end applications of ASV systems are usually the ones including computer log-in, telephone-based banking transactions, access to restricted buildings, personal identification, etc. which would be at high risk if the verification systems can be defrauded. More literature and work in professional mimics can be found in [**9**], [**12**], [**13**], [**14**].

Generally, the ASV systems are evaluated on zero-effort attacks. Zero-effort impostors are casual impostors where no effort is made to mimic or produce a speech as that of the enrolled speaker. With zero-effort impostors, current ASV systems can achieve very high accuracy and significantly low EER. It has been reported that with current techniques such as JFA [**10**] and PLDA [**11**] very low % EER is obtained for the ASV task. Hence, research has progressed in ASpR field in terms of overcoming the performance degradation due to microphone variability, intersession variability, speaker variability, recording conditions, etc. However, ASV systems should be robust to both zero-effort impostor trials and deliberate-effort spoofing attacks. The zero-effort case is an unrealistic scenario, as there is no advantage in mimicking a person without knowing anything about him or her. In a realistic scenario, an impostor has information about the target speaker. Thus, verification systems must be robust to both zero-effort attacks and deliberate-effort spoofing attacks as well. It must identify or discriminate between a natural speech from a true claimant and an impostor speech trying to mimic any target on its own or by utilizing available techniques of cut-paste, synthesis or voice conversion to sound like any of the intended target speaker.

1.4 Motivation for Spoof Detection Problem

Research in general spoof detection task is relatively well established with several competitive evaluations having been held for other biometric modalities such as face [**15**], fingerprint [**16**] and iris [**17**] recognition. In case of voice biometrics, the lack of availability of statistically meaningful standard datasets, protocol and metrics were initially a hindrance to study of spoofing and evaluating the performance of anti-

spoofing measures on a generalized platform. An initial attempt to create a base for standardization in the evaluation of countermeasures for spoofing was carried out with the organization of a Special Session at INTERSPEECH 2013 entitled, 'Spoofing and Countermeasures for Automatic Speaker Verification', wherein, mimic, replay, synthesis and voice conversion attacks were considered [**18**]. However, the various countermeasures proposed in [**18**] used prior knowledge of specific spoofing attack without any standard datasets, protocols or metric to measure and alleviate the possible threat of spoofing. Impersonation and replay attacks may be highly vulnerable when used as a spoof. However, they have their limitations in the context of developing countermeasures (which will be discussed in Chapter 2). Among the various spoofing methods, spoofs due to Synthetic Speech (SS) and Voice Converted Speech (VCS) are easily available and can be generated for any given text and for any speaker. With respect to this, recently the Spoofing and Anti-Spoofing (SAS) corpus has been developed providing a generalized dataset with Text-To-Speech (TTS) synthesis and voice conversion attacks for various spoofing algorithms on a large set of speakers [**19**]. Using a subset of the SAS database, very recently, the 'ASV spoof 2015 challenge' was organized as a special session of INTERSPEECH 2015 [**20**]. For this challenge, the task was to design an ASV-independent standalone detector that could classify natural and spoofed speech for both known and unknown attacks. The final results in terms of % EER were returned by the organizers of the challenge for both attack-dependent (i.e., known) and attack-independent (i.e., unknown) case. Thus, as shown in Figure 1.2, there exists a need for an independent or standalone detector for natural *vs.* spoofed speech prior (or post) to the ASV systems. In this thesis, we work towards developing suitable features to classify natural and spoofed speech and hence, can be used as countermeasures to alleviate possible spoofing attacks in voice biometrics.

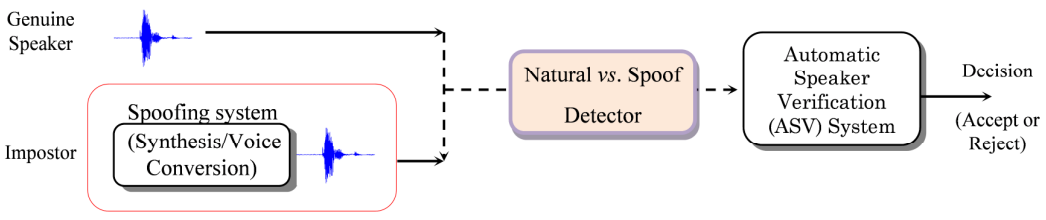


Figure 1.2: Spoofing on ASV system and the need for natural *vs.* spoofed speech detector.

At the challenge, for SS and VCS spoof detection, phase-based features (both Fourier transform and analytic or instantaneous) were used extensively. This was due to the fact that state-of-the-art SS and VCS generation techniques use vocoder which lacks phase information. These countermeasures gave almost *0.00 %* EER for known attacks. However, many of these approaches at times failed for unknown vocoder spoofs and in fact failed for unknown vocoder-independent spoofing attacks. Therefore, the research needs to be directed such that the countermeasures are effective in real-life scenarios, where the type of spoofing will *not* be known at all.

To detect natural *vs.* machine-generated speech, it is essential to use cues that are specific to the natural speech and absent in the machine-generated speech. The human speech mechanism has two main components, i.e., the vocal tract system and input to the vocal tract, i.e., excitation source. The acoustic speech output is a result of combination of a source of sound energy (e.g., the larynx) modulated by a transfer (filter) function determined by the shape of the vocal tract. This model is often referred to as the "source-filter theory of speech production" [**21**]. The source-filter theory describes speech production as a two-stage process involving the generation of a sound source, with its own spectral shape and spectral fine structure, which is then shaped or filtered by the resonant properties of the vocal tract system. Therefore, it is crucial to study the characteristics of the natural and spoofed speech signal from the excitation source and system point of view. The study of source and system characteristics separately assumes a linear speech production mechanism where the source and system can be independent. However, the actual speech production mechanism is a nonlinear phenomenon. Therefore, in this work, three basic aspects of speech, i.e., *excitation source*, *vocal tract system* (i.e., filter) and the *Source-Filter* (S-F) interaction or coupling information are explored. We believe that it is essential to study the independent contributions of the system-based and the source-based features for the SSD task. In addition, during natural speech production, neither *source* excitation nor vocal tract *system* alone is important, rather how they interact or couple is also essential which motivates for the use of S-F interaction features as well. Thus, as a foundation for the development of features, we use features that are derived from the understanding of the natural speech production mechanism and hence, these features will not be specific to the spoofed speech rather they are expected to be discriminative w.r.t natural *vs.* spoofed speech.

The threat of spoofing attacks has restricted the use of ASV systems for security applications like telephone banking, access to restricted areas/buildings, etc. Now that a generalized dataset is made publicly available [**20**], there has been a keen research interest in addressing this research issue. Simultaneously, it has been argued that low-technology spoofing attacks such as replay are more vulnerable and available even for intruders without speech processing knowledge. However, there was no standard dataset to evaluate the performance of the countermeasures for replay until the recent ASV spoof 2017 challenge [**22**]. Moreover, replay attack is feasible with text-dependent systems where the keyword is known. On the other hand, tools are readily available to generate synthetic or converted speech without requiring much higher levels of expertise. Thus, exploring suitable countermeasures for spoof detection of SS and VCS is also highly essential. Detecting spoofed speech not only aids to have secure ASV systems, but also, has various other applications as discussed in the next sub-Section.

1.5 Applications of Spoofed Speech Detection (SSD)

The problem of detecting spoofed speech is highly essential and needs to be addressed. Few of the applications of SSD are:

- Spoof detection is necessary for the security of ASV systems. Reliable ASV systems are essential for telephone banking, personal identification and computer logins, etc.
- It can be used for *liveliness* detection in speaker forensics, where it is essential to know if the speech recording is from the actual suspect or an attempt is made to indulge the suspect by making an unauthorized access.
- Based on the countermeasures proposed, lacunas or artifacts in the spoofed speech can be studied to investigate reasons of quality degradation in the spoofed speech.
- Based on the above, the countermeasures could be used as objective measures for the evaluation of speech synthesis and voice conversion systems so as to investigate how much the synthetic speech is close to natural speech or how much the voice converted speech is similar to the target speaker's speech.
- Depending on the countermeasures, the *differences* between actual human speech production model and the simplified model can be studied. These can

then be used in improving SS and VCS generation algorithms for improved naturalness and speaker similarity. However, improving naturalness in the SS and VCS may also affect the performance of the proposed spoof detection system, which further needs to be improved.

The detection of SS speech is essential because any random text can be generated for any speaker and in VCS spoof, any speaker can be targeted (i.e., even from male-to-female and vice-versa is possible).

1.6 Contributions from the Thesis

The main focus or approach of the thesis is the development of suitable features for SSD task. In this thesis, three basic aspects of speech production mechanism, i.e., *excitation source*, *vocal tract system* (i.e., filter) and the *S-F* interaction features are explored to design countermeasures for SSD task. The brief details about the various features proposed are shown in Figure 1.3 (where the dotted blocks indicate the features used in the literature for SSD task).

1.6.1 Source-based Features

For excitation source features, we propose the following feature sets which when combined with the system-based features, decreased the % EER of the SSD system.

- **Fundamental frequency (F_0) contour and Strength of Excitation (*SoE*) features**: When the vocal folds vibrate, there exists a correlation between the F_0 contour and *SoE* at the glottal excitation source and at the speech signal. This correlation is found to be more for natural speech than machine-generated speech. Moreover, natural speech has variations within the F_0 contour and *SoE* depending on the speaker and speech characteristics which may not be the case for spoofed speech.
- **Prediction-based features:** Here, we propose the use of Linear Prediction (LP), Long-Term Prediction (LTP) and Non-Linear Prediction (NLP) features. The idea is that the spoofed speech is too easy to predict if a simplified acoustic model generates it and it is too difficult to predict if there are artifacts present in the speech signal. The nonlinearity in speech is an attribute of natural speech production mechanism and hence, LP-NLP combination provided better discriminative features as compared to existing LP-LTP approach.

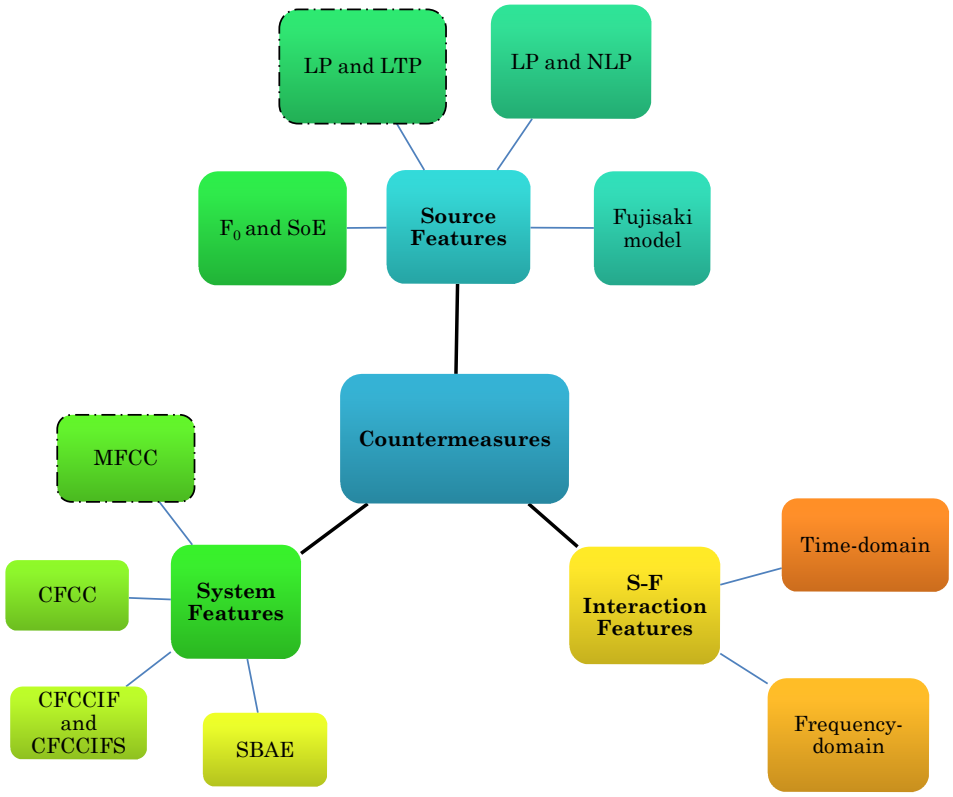


Figure 1.3: Classification tree of various features used as countermeasures (dotted boxes indicates approaches used in literature and rest indicates the contributions in the thesis).

- **Prosodic features derived from the Fujisaki model:** Humans use the prosodic cues from speech to identify naturalness in the speech signal. Thus, the accent and phrase parameters from Fujisaki model assists in using the higher-level information to distinguish the two classes. Hence, Fujisaki model has been explored for adding prosodic features in the synthesized speech to improve the naturalness of TTS generated voice [**23**]. However, in this thesis work, we attempt to utilize it for a counter problem (in particular, to detect lack of prosodic features in the spoofed speech).

1.6.2 System-based Features

For system features, we explore the underlying idea that the human ear processes speech in subbands (due to the signal processing abstraction of the cochlea, i.e., vibration of basilar membrane in a specific region for a specific tone). Moreover, the human speech production system does not produce speech in a frame-by-frame pattern (rather in a *continuum* which implicitly captures the naturalness in speech

production mechanism) while feature extraction in SS and VCS is generally at frame-level. Hence, dynamic variations across frames are significant for SSD task. With respect to this, cochlear-based features and Deep Neural Network (DNN)-based features are proposed:

- **Subband Envelope and Instantaneous Frequency features:** Instead of the MFCC feature set, the cochlear filter representation, i.e., Cochlear Filter Cepstral Coefficients (CFCC) features which mimic the auditory system more efficiently are used and also modified to derive new features set for SSD task. In particular, the envelope of the output of the cochlear filter is combined with the average IF to give CFCCIF features. The basic idea is that the envelope of each output of the cochlear filter and its analytic phase are important features used by the auditory levels for speech perception (Chap. 8, pp. 403 [**21**]). Moreover, to capture transient information or the variation across frames, the derivative operation is used. The resultant feature detected even non-vocoder spoofed speech (to a certain extent), and it performed best on an average for known and unknown attacks at the ASV spoof 2015 challenge.
- **Subband Autoencoder (SBAE)-based features:** A new architecture of Autoencoder (AE) is explored that embeds the subband processing in Human Auditory System (HAS). This data-driven approach is used to learn features from the speech spectrum. The SBAE features are found to capture more dynamic information across the frames of the speech spectrum. As a result, the vocoder-independent spoofs were detected well.

1.6.3 Source-Filter (S-F) Interaction-based Features

Next, we explore the fact that the nonlinear S-F interaction is an attribute of the natural speech production mechanism and it is highly complex to build or mimic such nonlinear interaction while synthesizing speech artificially. Based on this, we propose using the following features for the SSD task:

- **Shape and residual energy-based features in the time-domain:** The L^2 norm of residual signal ($g_r(t)$) between the glottal flow derivative waveform ($\dot{g}(t)$) and its fitted Liljencrants-Fant (LF) model ($g_c(t)$) along with the shape features from the fitted model, in the closed, open and return phases of the

glottis are considered as features. With fewer feature dimensions, not only did the features work well for vocoder-based spoof speech, but also, the features performed well in noisy and signal degradation conditions.

- **Residual energy-based features in the frequency-domain:** In the frequency-domain, the Mel representation of residual $g_r(t)$ and the residue or difference of the spectrogram (as well as the Mel-warped spectrogram) of the estimated $\dot{g}(t)$ and $g_c(t)$ is found to have complementary information than time-domain features for the SSD task.

Finally, all the features have been evaluated under unknown attacks such as speech generated by various algorithms in the Blizzard Challenge 2012 database [**24**]. Similar evaluation has also been carried out on Hindi and Gujarati language using the Blizzard Challenge 2014 database with both vocoder-dependent and vocoder-independent synthetic speech [**25**]. This helps to evaluate the performance of the features for completely unknown attacks and also for channel mismatch conditions.

1.7 Organization of the Thesis

The organization of the thesis is shown in Figure 1.4 and is discussed in detail below:

Chapter 2 discusses the literature survey on spoofing attacks for voice biometrics. Various spoofing attacks are discussed with special emphasis or focus on Hidden Markov Model (HMM)-based speech synthesis and voice conversion attacks. A detailed review about the methods identifying the vulnerability of spoofing attacks on ASV is presented. This is followed by discussing of various countermeasures existing in the literature with ASV systems and without ASV systems (i.e., as a standalone spoof detector). The several issues with the stand-alone detectors are also briefly discussed.

Chapter 3 deals with the spoofing techniques and the general architecture of the spoof detection system. The speech synthesis and voice conversion techniques are discussed in detail. Next, in the spoofed speech detection architecture, the databases used for the study, classification system and the performance measures are discussed. With respect to the database, details about the spoofing algorithms in the ASV spoof database and Blizzard challenge data are discussed. The details about the TTS building procedure for Gujarati language both using the Unit Selection Synthesis (USS) and HMM-based TTS Synthesis System (HTS) framework are

provided as part of the Blizzard challenge database. The brief details about Gaussian Mixture Model (GMM)-based classification system and the performance measures in terms of Equal Error Rate (EER) and Detection Error Trade-off (DET) curve are discussed.

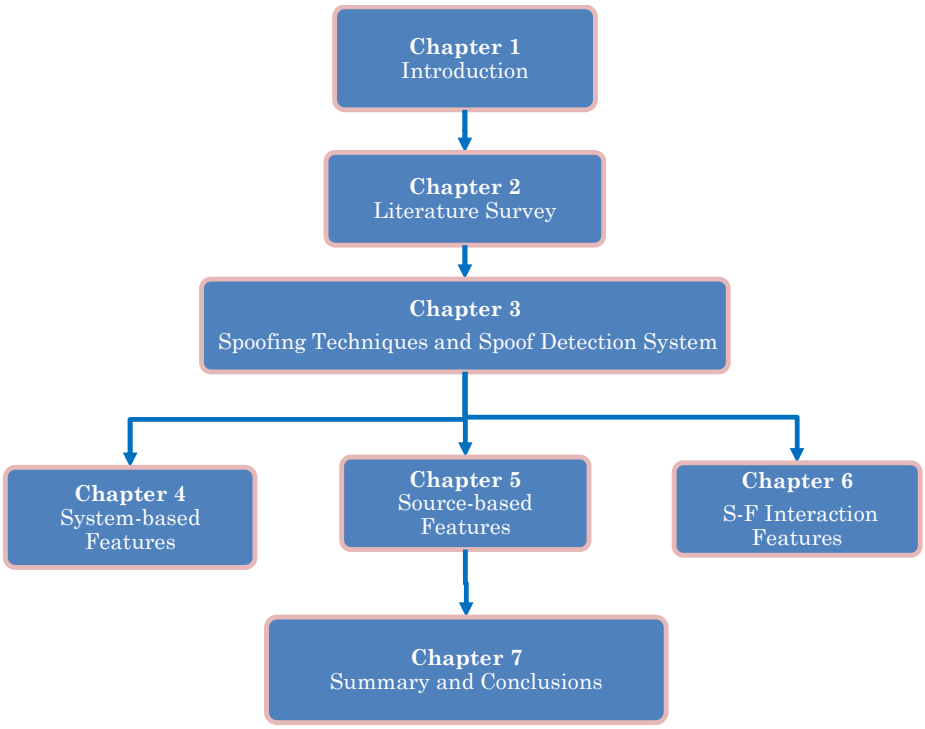


Figure 1.4: Organization of the thesis.

Chapter 4 discusses the various system-level features for SSD task. This includes the MFCC features, CFCC features and the proposed CFCCIF and CFCCIFS feature sets. Furthermore, a data-driven approach is used to learn features from the speech spectrum using SBAE. The experimental results of all the features are presented and discussed.

Chapter 5 discusses the various source features used in the study. This includes F_0, *SoE* and their dynamic variations. The algorithms used to extract F_0 and *SoE* from the speech are also discussed. In addition, features derived from various prediction techniques such as LP, LTP and NLP for SSD task are presented. Furthermore, the Fujisaki model is studied in detail and the prosodic differences between the natural and synthetic speech are analyzed.

Chapter 6 discusses several S-F interaction features that are peculiar to the natural speech signal. The procedure for estimating the glottal excitation source and the LF-model are discussed. Followed by this is the development of various time-domain and frequency-domain features and their performance for the SSD task.

Chapter 7 concludes and summarizes the work done in the thesis. The contributions in the thesis are presented. The Chapter also discusses the applications, limitations of the present work and future research directions for the task of anti-spoofing presented in the thesis.

1.8 Chapter Summary

This Chapter gave an outline of the basic ASV system and introduced the problem of spoofing. The motivation and need of anti-spoofing measures are discussed and an overview of the countermeasures proposed in the thesis for spoofed speech detection is provided. The next Chapter discusses in detail the development of various spoofing types and the countermeasures proposed in the literature to overcome the spoofing attacks. The limitations of the measures existing in the literature along with current research issues in this area are discussed. In the next Chapter, a selected chronological literature search in the proposed area of study is presented to address the problem of spoofing and identify new directions for research.

Chapter 2.
Literature Survey

2.1 Introduction

This Chapter discusses the literature survey on the Spoofed Speech Detection (SSD) problem. Various possible spoofing attacks are discussed along with their pros and cons to be considered when dealing with the SSD task. Evolution of the vulnerability of Automatic Speaker Verification (ASV) system to speech synthesis and voice conversion attacks is presented. The anti-spoofing measures existing in the literature both with and without ASV system are discussed. This Chapter brings out briefly the research issues in the current approaches for the SSD task, majority of which will be addressed in this thesis work.

2.2 Spoofing Attacks

Impostor speech can be generated by humans themselves or artificially by using computers. Impostor due to humans is a case of impersonation (or that of mimicking) or that of identical twins. On the other hand, impostor attack through machines can be due to replay, speech synthesis and voice conversion techniques as well. A brief discussion on these attacks in voice biometrics is discussed in this sub-Section and shown in Figure 2.1.

2.2.1 Mimics

Mimics can be due to *physiological* characteristics that are observed in identical twins or due to *behavior* or learned characteristics or features such as professional mimics. An ASV system must have high mimic resistance capability to distinguish between a mimic and a natural speech from a genuine speaker. Spoof detection by mimic is a difficult task especially in the case of identical twins, as the twins are likely to have nearly similar (if not identical) vocal tract shape and size. Therefore, it will be difficult to distinguish the two voices. On the other hand, for professional mimics, the prosodic characteristics can be mimicked. However, the actual vocal tract

structure of source speaker cannot be made identical to that of target speaker [**14**]. Mimicking is a straightforward spoof that can hamper the security of ASV systems without prior knowledge of any speech processing technology or computer-aided techniques. However, several reasons exist, that refrain their use as a spoofing attack and related research works based on physiological characteristics. Firstly, in the case of mimics, to create a spoof, a twin is needed and this is a very rare case. Secondly, in the case of impersonation, some impostor speakers have a natural ability to be confused with or sound like other speakers. Likewise, few target speakers may be easily impersonated than the others. Thus, spoofing and anti-spoofing are dependent on the selection of target speaker's characteristics. The survey reported in [**26**], reports inconsistency in the findings of the vulnerability of mimic attacks on various ASV systems. This may be due to the reason that the studies are carried out on small datasets with different speakers. Hence, in this thesis, we do not consider mimics for developing anti-spoofing measures.

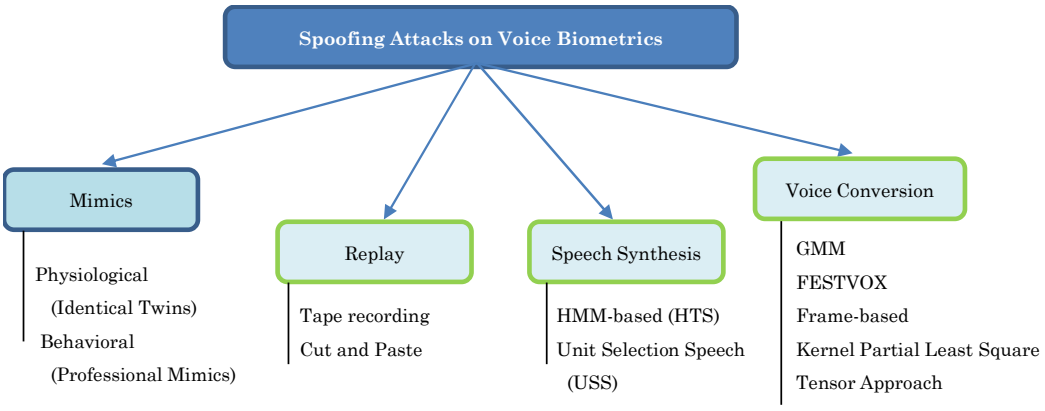


Figure 2.1: Types of spoofing attacks on voice biometrics.

2.2.2 Replay

Replay is a type of spoofing attack where an impostor tries to use pre-recorded speech samples that are collected from selected target speaker to access his or her biometric system illegally. The recordings can be over a period of time and cut and concatenated (pasted) as per the required phrase or segment that is the intended password or a keyword. This kind of spoof is highly unsafe because if the speech recordings are of high quality, then they will be exactly similar to the real speech signal. In addition, due to the availability of high-quality and inexpensive recording devices, such as smartphones, replay attack is very accessible. Moreover, the

attacker does not need any advanced technical knowledge or the expertise on computer-aided techniques. This type of spoof is vulnerable to text-dependent ASV systems [**26**]. However, the attacker must have proximity to the target to get the recording which is not always the case. Approaches have been made to detect replay speech played back through handheld devices [**27**]. Due to the severity of this spoofing attack, research in this direction is carried out with visual aids as well [**28**]. In considering speech alone, there are various assumptions that can be made while detecting replay attacks [**28**]. In this work, we do not consider this extreme case of a spoofing attack. We use the fact that computer-aided technologies are also equally unsafe for the security of ASV systems. Later, we consider the case of unit-selection synthesis attack which is also a cut-paste computer generated approach (where speech sound units are selected by minimizing the cost of joining units).

2.2.3 Speech Synthesis

Speech synthesis technology is also referred to as Text-to-Speech (TTS) synthesis technique. The TTS technology aims at generating intelligible and natural-sounding machine-generated speech for a given input text. The TTS technology has a wide range of applications, i.e., development of tools for the visually impaired and as communication aids for the speech impaired, e-book reading, for in-car navigation systems, storytelling application, singing speech synthesizers, spoken dialogue systems, and in speech-to-speech translation systems. General speech synthesis systems have two main components, i.e., text analysis (front-end) and the speech waveform generation (back-end). In the text analysis component, the input text is converted into a linguistic form that consists of components such as phonemes. Next, in the waveform generation component, the speech waveforms are generated from the produced linguistic representation. Considering the waveform generation process in detail, there are four major approaches that were developed over the decades. Initially, in the 1970's, the formant synthesizers were developed that used *handcrafted* acoustic rules such as vocal tract area functions, articulatory parameters, formant resonances, etc. for each phoneme segment [**29**]. Later in the 1980s, instead of manually crafting rules, the use of pre-recorded speech concatenation approach was used. This started with a small database of phoneme units called 'diphones' and these are concatenated according to the given phoneme sequence through signal processing techniques [**30**]- [**31**]. Later in the 1990s, larger

speech databases were collected and appropriate speech sound units were selected for concatenation. These units were selected based on linguistic contexts, intonation, syllable position, the first and last position of the word or the phrase, accent, etc. This approach is referred to as Unit Selection Synthesis (USS) and is known to be more natural-sounding speech than speech generated by diphone-based synthesis systems [**32**]. In the late 1990s, the approach of Statistical Parametric Speech Synthesis (SPSS) was introduced [**33**]- [**34**]. In this approach, the Hidden Markov Model (HMM) is used to model several acoustic parameters from speech. HMMs take into consideration the knowledge of phoneme sequences and various contexts of the linguistic specification (i.e., sequential information using a statistical model that describes the production of speech signal). The acoustic parameters that are generated from HMMs and selected based on the linguistic specification are given to a vocoder (which is a simplified speech production model represented by vocal tract and excitation source parameters) to generate required speech waveform. The HMM-based speech synthesizers can not only learn models from relatively less speaker-specific data, but also possibly adapt to the background models for new speakers [**35**]- [**36**]. The initial three approaches, formant synthesis, diphone synthesis and unit-selection synthesis are not much likely to be effective in ASV spoofing. This is because formant synthesis does not offer synthesis of speaker-specific formant characteristics whereas di-phone or unit-selection approach requires a large speaker-specific database. The database must cover all the diphones and also several instances of the diphone resulting in a large amount of speaker-specific data along with appropriate labeling of the speech data (which is a very laborious task and time-consuming even for a single speaker). Furthermore, to develop the statistically meaningful database for SSD task, a large number of voices (i.e., from different speakers) are required, which was possible in HTS and not feasible in case of USS synthesis (due to the tedious task of labeling of speech sound units). Thus, the use of HMM-based speech synthesis was generally preferred as a possible candidate for spoofing over other TTS techniques. However, with the development of open-source speech synthesis systems like MARY (Modular Architecture for Research on speech sYnthesis [**37**]) it has been possible to develop TTS systems for large number of speakers and hence, the use of USS-based spoofing attacks can really fool a biometric system. It should be noted that USS-based spoofing attacks are highly speaker-specific and if used as a spoof can really fool a biometric system.

2.2.4 Voice Conversion

Voice conversion technique aims to change the speech of a given *source* speaker so that it may sound-like or resemble in some sense that of another speaker called as *target* speaker [**38**]. Application of voice conversion includes voice editing and dubbing (to preserve the voice of speakers), medical applications (such as voice restoration and interfaces for speech pathologies). Unlike TTS, which takes the text as input, the voice conversion system takes natural speech signal as an input signal. Typically, voice conversion technique involves both timbre and prosody conversion. Timbre relates to spectral envelope of the speech signal. Thus, in voice conversion, spectral mapping from source to target speaker is to be carried out. On the other hand, prosody conversion relates to converting prosodic features, such as the fundamental frequency (F_0) and duration. Considering spectral mapping, the approaches include vector quantization, statistical parametric approaches, frequency warping, unit-selection and neural network-based techniques. The initial approach included spectral mapping based on vector quantization that included a mapping codebook from source-to-target feature pairs [**39**]- [**40**]. The popular statistical parametric approach in the literature includes using Gaussian Mixture Model (GMM)-based voice conversion. The GMMs can be trained on source features [**41**] or the joint density of both source and target speakers' features [**42**]. The data is converted by a function that is a weighted sum of local regression functions. Next, the frequency warping method includes developing a warping function between the source and target speech spectra. The frequency warped source spectrum on combining with GMM-based converted spectrum is known to reduce the effect of *over-smoothing* [**43**]. This technique gives a good quality voice, however, the speaker similarity may not be as close to the target due to difficulty in preserving *shape* and the *-3 dB bandwidth* of the formants. To overcome this issue, many warping functions have also been introduced in the literature [**44**]. As in TTS synthesis, unit-selection in voice conversion is also studied to directly utilize the target speaker's speech segments for conversion [**45**]- [**46**]. Approaches based on neural networks also exists to model the nonlinear relationships between the source and the target speakers [**47**]- [**48**]. In addition to spectral information, the speech prosody information (which is a medium-level or speaker-specific feature) is also important to identify the speakers. The F_0, intonation and duration are generally used as prosodic

features. Various approaches exist to convert a source speaker's F_0 contour to that of the target speaker [**49**]- [**50**] whereas phoneme or syllable duration conversion approaches were reported in [**51**]- [**52**]. Voice conversion technology is likely to be effective in attacking ASV systems. Spectral mapping techniques shift an impostor's spectral characteristics to match that of a specific target speaker and hence, present a threat to ASV systems which generally use spectral features only. Meanwhile, prosody conversion can manipulate an attacker's prosodic characteristics to mimic those of a target speaker and thus, they present a risk to those ASV systems which use prosodic features for ASV task [**26**].

Therefore, various algorithms exist to generate both SS and VCS. Hence, the anti-spoofing measures need to be tested on various SS and VCS algorithms to check the performance or anti-spoofing capabilities of the countermeasures. With respect to this, to provide a large statistically meaningful database for evaluating the countermeasures, the ASV spoof 2015 database was introduced with spoofed speeches from three speech synthesis and seven voice conversion generation algorithms [**20**]. The details about the database and the techniques of spoofed speech generation are discussed in Chapter 3.

To get initial insights to the differences between natural and spoofed speech, we observe their Short-Time Fourier Transform (STFT) representation as in Figure 2.2. The STFT of the speech is taken as follows,

$$S(m,\omega) = \sum_{n=-\infty}^{\infty} s(n)w(n-m)e^{-j\omega n}, \quad \textbf{(2.1)}$$

where $s(n)$ is the speech signal and $w(n)$ is the analysis Hanning window. Next, the power spectrum is computed. Both the narrowband and wideband spectrograms are observed with a frame window of *25* ms and *5* ms, respectively. In Figure 2.2, for the same text, Panel I, Panel II, Panel III and Panel IV corresponds to natural speech, vocoder-based SS, vocoder-based VCS and USS-based speech, respectively. It is observed from the Figure 2.2, that it is very difficult to distinguish between the natural and spoof speeches from the narrowband spectrogram. Even for the wideband spectrogram, not many differences are visible, except for the very low frequencies energies present in the silence regions at the start and end of the utterance. These energies may be due to the microphone or the recording device and

hence, they are absent in SS and VCS. In the case of USS-based speech, such low frequency energies are present due to the fact that USS-based speech is made from the concatenation of natural pre-recorded speech sound units. However, this difference may not be useful for classification of natural and spoofed speech. This is because, in real case scenarios, the spoofed speech will be played via some recording device to the biometric system. Hence, this difference may not be prevalent between the natural and spoofed speech. Thus, spectrographic analysis does not show significant differences for natural *vs.* spoofed speech and hence, there is a need of developing anti-spoofing measures to distinguish between the two speech signals.

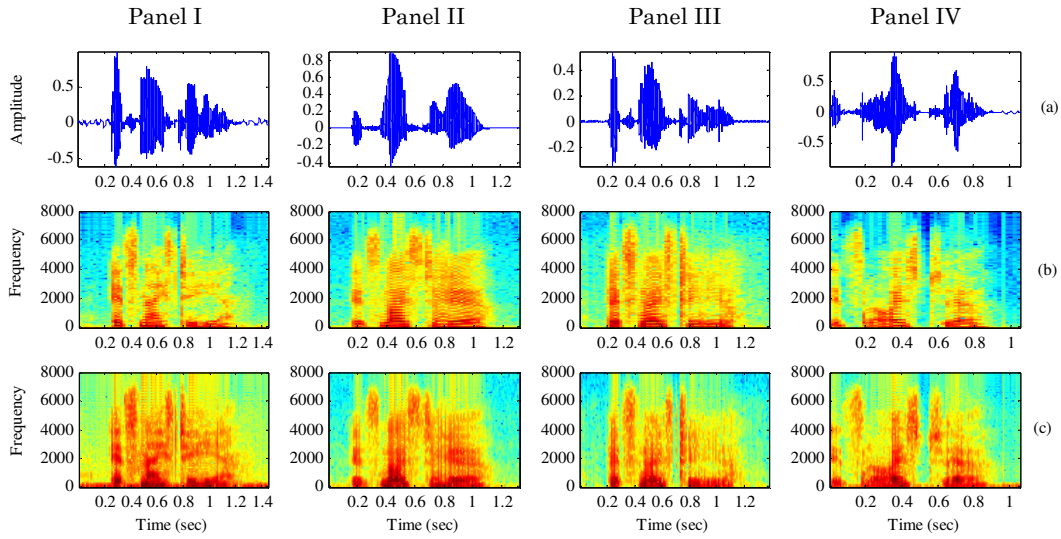


Figure 2.2: The spectrographic analysis of natural *vs.* spoofed speech: (a) speech signal, (b) narrowband spectrogram and (c) wideband spectrogram. Panel I: Natural speech, Panel II: vocoder-based SS, Panel III: vocoder-based VCS and Panel IV: USS-based speech.

2.3 Vulnerability of ASV Systems to Spoofed Speech

During the 1990's, the computer-aided voice conversion techniques were more popular than speech synthesis due to the fact that the quality of the synthetic speech was not good enough and also not adapted to any arbitrary speaker. Hence, earlier work includes demonstrating the vulnerability of ASV systems against VCS spoof rather than SS. In 1998, the first study considering attack due to converted speech was reported for HMM-based ASV system [**53**]. In this work, an analysis and synthesis technique of Harmonic plus Noise Model (HNM) is used to obtain the speaker transformation. It was observed that the quality of HMM-based transformation was more efficient than GMM-based voice transformation. With an

HMM-based transformation attack, the baseline EER of ASV increased from *4.19 %* to *23.09 %*. Followed by this, the sensitivity of GMM-based ASV to VCS spoof was studied [**54**]. For the ASV system trained on *138* speakers of the YOHO database, the % EER increased from *1.45 %* to *86 %*. If the EER threshold was varied to obtain a False Rejection Rate (FRR) of *25 %*, the False Acceptance Rate (FRR) was still *34.6 %*, which was sufficiently high. Simultaneously, the improvement in HMM-based synthesis techniques [**55**] and initial work of using speaker adaptation techniques [**56**], initiated the need to consider the effect of SS as an impostor to ASV system.

In 1999, Masuko *et al.* carried the first work to recognize the vulnerability of text-prompted HMM-based ASV systems against SS impostor attacks [**57**]. The HMM-based synthesis with speaker adaptation technique to convert into a target's voice was used to spoof the ASV system [**55**], [**58**]. Baseline ASV with *0.00 %* EER reached above *70 %* EER after spoofing. Simultaneously, Lindberg *et al.*, [**59**] studied the effect on FAR for various spoofing types for a male and a female speaker. An ASV system with a baseline EER of *5.6 %* under human impostor, after spoofing with concatenated speech, re-synthesized speech and di-phone-based synthesis systems gave a FAR of *94.75 %*, *8.75 %* and *38.9 %*, respectively. Thus, in addition to HMM-based SS spoofs, speech generated by concatenation of speech sound units is more prone to affect the ASV systems. Thus, USS or concatenation approach is also very effective as an impostor attack and thus, one of the most challenging spoofs to detect. However, initially, speaker adaptation with USS was not well researched to use it as a generalized spoofing attack. Due to the recent development of HMM-based speech synthesis [**35**]- [**60**] and voice conversion techniques [**13**]- [**41**] along with its generalization for any speaker using adapted HMM-based systems [**36**], several studies demonstrated the vulnerability of the ASV systems to SS and VCS attacks. In fact, it has been observed that ASV systems are more vulnerable to VCS attacks than SS attacks. The VCS spoof has shown to drastically increase the % EER for GMM-based ASV systems [**61**]- [**62**]. For Joint Factor Analysis (JFA)-based ASV systems [**63**] and *i*-vector-based ASV systems [**64**], the FAR has increased more than *5*-folds. A detailed summary of effect for SS and VCS attacks on ASV system can be found in [**26**].

2.4 ASV-independent Spoof Detection

The design of stand-alone detector is considered as an independent research problem (for example, as in the ASV spoof 2015 challenge [**20**]). The basic task is to detect natural and spoofed speech (generated by any spoofing algorithm, i.e., either speech synthesis or voice conversion). In this Section, we emphasize and discuss the countermeasures that have been proposed for the detection task (both for clean and signal degradation conditions). The countermeasures are basically divided into two types based on their design, i.e., for known attacks and unknown attacks. A brief representation of the entire literature search is also shown in Figure 2.3.

2.4.1 Countermeasures for Known Attacks

Table 2.1 shows the summary of the various countermeasures, along with details of the database used while testing, type of spoof, classifier details and detection accuracy or EER. In [**65**], the authors propose using time stability and pattern of the F_0 contour to discriminate between natural and HMM-based synthetic speech. It was observed that for *100* utterances of a male speaker, the time stability was more for SS speech than natural speech signal. Next, measures such as, peak, lower half, the upper half and half bandwidth are extracted from the F_0 pattern. A dynamic programming distance factor is used between the enrolled natural speech of the ASV system and the tested SS or natural speech. This distance is more for the SS and hence, can be useful for discriminating SS and natural speech. However, the experiments of time-stability and pitch pattern analysis were done on less number of speakers, i.e., *1* male (*100* utterances) and *5* male (*20* utterances). On the similar lines of using statistics derived from F_0, in [**66**], image processing approach was used to extract Mean Pitch Stability (MPS) and MPS range (MPSr) features from F_0 and an additional jitter feature was utilized. A GMM-based classifier was trained on the National Institute of Standards and Technology (NIST) 2002 and Blizzard Challenge 2008-2011/Festival built-in voices for natural and spoofed utterances. The evaluation was carried out by using natural and SS from the Switchboard and Wall Street Journal (WSJ) corpus, respectively. The accuracy of detecting SS was observed to be *96 %* which was better than *88 %* obtained by Relative Phase Shift (RPS) used by the authors in [**67**]. The F_0 patterns generated for SPSS approach tend to be over-smoothed and USS approach exhibited jumps at the concatenation points.

Table 2.1: A summary of the various countermeasures used for SS and VCS spoof detection on known attacks

Study	Evaluation Corpus	Spoof Type	Features	Classifier	Accuracy (in %)	EER (in %)
Ogihara, *et al.* (2005) [**65**]	1 (Male) 100 utterances	HMM-SS	Time Stability	-	90	-
			Pitch Pattern	-	99	-
De Leon, *et al.* (2012) [**66**]	Switchboard/ WSJ/RM	HMM-SS	MPS, MPSr, Jitter	GMM	96	-
Z. Wu, *et al.* (2012) [**68**]	NIST 2006	GMM-VCS	MFCC	512-GMM	-	16.80
		GMM-VCS	Cos-phase		-	6.60
		GMM-VCS	MGDF-phase		-	9.13
		USS-VCS	MFCC		-	15.35
		USS-VCS	Cos-phase		-	3.93
		USS-VCS	MGDF-phase		-	4.60
Z. Wu, *et al.* (2013) [**69**]	WSJ0+WSJ1	VCS	MFCC	512-GMM	-	10.98
			MGDF-phase		-	1.25
			MM		-	19.29
			PM		-	13.71
J. Sanchez, *et al.* (2015) [**70**]	WSJ	HMM-SS	MFCC	512-GMM	-	16.03
			RPS			1.22

Note: '-' indicates that the details are not available

Next, an approach to detecting GMM-based converted speech and unit-selection-based converted speech data is used in [**68**]. The NIST 2006 SRE corpus is used with *100* sessions of speech to train the system and *1500* sessions of natural speech, *1000* sessions of GMM-VCS, *1000* sessions of USS-VCS for testing. Cosine normalization of phase spectrum (Cos-phase), i.e., the Discrete Cosine Transform (DCT) of the cosine of unwrapped phase spectrum is used as features in addition to DCT of the frequency derivative of the Modified Group Delay Function (MGDF). An average EER of *5.265 %* and *6.865 %* using the Cos-phase and MGDF-phase features is obtained, respectively. The EER obtained by state-of-the-art MFCC features is relatively higher than the phase-based features. One of the reasons is that both the GMM-VCS and USS-VCS spoof introduces phase artifacts in phase spectrum which is captured by Cos-phase and MGDF-phase features. On the other hand, Fourier transform phase features are completely ignored in MFCC. A similar approach to using phase-based features was made in [**69**], using a supervector of Magnitude Modulation (MM) and phase modulation (PM) features. The MM and PM features were derived from *50* frames (*1* frame = *25 ms*) window of the power spectrum of speech and the MGDF-phase spectrogram, respectively. Thus, these features constitute the long-term features as compared to the MFCC and MGDF-phase features which are short-term features extracted at *25 ms* frame length. As seen from Table 2.1, the MGDF-phase features gave very low *1.25 %* EER. When these

features were fused at score-level with each other, it was observed that the combination of MGDF-phase function with MM and PM gave % EER of *0.98* and *0.89*, respectively. This shows that long-term temporal features have complementary information than short-term spectral features which was observed by their score-level fusion. The work aimed to prove that artifacts are introduced during the frame-by-frame operation in speech analysis and synthesis process which can be explored for SSD task. The study reported in [**70**], continued comprehensive evaluation using RPS and MFCC to develop a standalone SS detector. The RPS and MFCC features were evaluated on different conditions such as cross-vocoder testing and multi-vocoder testing. A 3-vocoder model was proposed that obtained *1.2* % and *16.03* % EER on the test set with RPS and MFCC features, respectively. Thus, the SS were easily detected by RPS than using the MFCC feature sets. To consider the unknown case, the features were tested on the Blizzard Challenge 2011 and Blizzard Challenge 2012 databases.

2.4.2 Countermeasures for Unknown Attacks

For the unknown case, the majority of the work started with the development of Spoofing and Anti-Spoofing (SAS) database and the ASV spoof 2015 challenge dataset [**19**]- [**20**]. Details of the challenge database are given in the next Chapter. At the challenge, a generalized and statistically meaningful dataset was provided and the results in % EER on the evaluation data were returned by the organizers. Table 2.2 shows the summary of the various countermeasures used for unknown attacks on the ASV spoof challenge database. As *S10* is a vocoder-independent spoof, its % EER is reported separately. The effect of the individual attacks on % EER and % FAR of ASV systems are reported in [**19**].

In [**71**], the authors use supervectors derived from the magnitude and phase information in the frequency-domain to detect spoofed speech. That is, they use the Local Binary Patterns (LBP) features jointly with the DCT of MGD-features and Cos-phase features. It is obtained that MGD-features and Cos-phase features, when fused with previously proposed LBP features, an EER of *0.058* %, is obtained on the development set and the details of the results on evaluation set are shown in Table 2.2. The results are quite well for vocoder-based spoofs which is not the case for the vocoder-independent *S10* spoof. This is due to the fact that no vocoder is used in an

S10 spoof, which does not bring any texture changes with LBP features and also the phase information is preserved to a great extent. Similarly, in [**72**], several features including Log-Magnitude Spectrum (LMS), residual LMS, Group Delay (GD), MGD, Instantaneous Frequency (IF), Baseband Phase Difference (BPD) and Pitch Synchronous Phase (PSP) were all fused at score-level to achieve the best EER of almost *0.00 %* on *S1-S9* spoof and due to similar reasons explained above, *S10* spoof remained almost undetected with *26.1 %* EER. In [**73**], in addition to MGD features, relative phase information extracted from the Fourier spectrum of speech is used to detect human and spoofed speech. The EER reduced from *1.74 %* for MFCC to *0.83 %* for MGD and *0.013 %* for relative phase alone. The *%* EER reported for *S10* spoof was as high as *37.068 %*. Similarly in [**74**], the RPS was used as features where the EER of unknown attacks was very high with *8.883 %*. The RPS features along with MGD phase was again explored in [**75**] where it was considered for a generalized spoofing scenario by testing the RPS feature set on the Blizzard 2012 Database.

In addition to phase-based features, Linear Prediction (LP)-based features were also explored [**76**]. The basic idea behind using prediction-based features is that spoofed speech is either easily predicted (if generated by a simple acoustic model) or either very difficult to predict (due to artifacts in the speech signal as in concatenated speech synthesis). Various features were derived based the LP residual and the Long-Term Prediction (LTP) residual. This approach was a novel way of considering spoofing attack detection. However, various issues were observed and the average *%* EER reported was around *11.6 %*. In addition, in [**77**], LP residual was used giving *8.9 %* EER on unknown attacks. However, the performance was not better than other phase-based approaches as shown in Table 2.2.

In addition to using various features for spoof detection, the use of various pattern classifiers was also explored. This includes *i*-vector-based systems [**78**], Deep Neural Networks (DNN)-based representation [**79**], use of DNN and Support Vector Machine (SVM) classifier [**80**] for spoof detection. In [**79**], a supervised DNN was trained using filterbank features on the training data of the challenge database. The combination of several features such as MFCCs, Mel Cepstral Coefficients (MCCs), Band Aperiodicity (BAP) and F_0 was used in this work. This approach was able to achieve an average EER of about *0.058 %* on known attacks and *5 %* on unknown attacks. In [**80**], DNN-based classifiers were used for spoof detection using LFCCs

Table 2.2: A summary of the various countermeasures used for SS and VCS spoof detection on the ASV spoof 2015 challenge database with the details of the feature sets used, classifiers and % EER for the known, unknown attacks and *S10* spoof on the evaluation set

Study	Feature Sets	Classifier	EER (%)			
			Known	Unknown	Average	S10
Liu, *et al.* (2015) [**71**]	Supervectors from MGD, Cos-phase. Fused with LBP features	GMM SVM	0.358	6.078	3.218	28.58
Xiao, *et al.* (2015) [**72**]	LMS, residual LMS, GD, MGD, IF, BPD, PSP	GMM	0.003	5.231	2.617	26.1
Wang, *et al.* (2015) [**73**]	Relative Phase Information	GMM	0.005	7.447	3.726	37.07
Sanchez, *et al.* (2015) [**74**]	RPS	GMM	0.21	8.883	4.547	40.00
A. Janicki [**76**]	LP and LTP	Logistic	6.1	17.1	11.616	-
Alam, *et al.* (2015) [**77**]	Cos-phase, MGD, Product Spectrum, LP residual	GMM	0.041	5.347	2.694	26.392
Patel, *et. al* (2015) [**81**]	MFCC+ CFCCIF	GMM	0.408	2.013	1.211	8.49
Weng, *et al.* (2015) [**78**]	*i*-vector framework (MFCC, MFCC-PPP (phoneme posterior probability)	KNN PLDA SVM	0.405	6.247	3.326	29.66
Chen, et. al (2015) [**79**]	Residual Log Magnitude Spectrum (RLMS), GD	DNN GMM	0.058	4.998	2.528	22
Villalba, *et al.* (2015) [**80**]	Fusion DNN (Spectrum+ RPS)	DNN	0.025	8.168	4.097	40.71
Novoselov, *et al.* (2015) [**82**]	MFCC, MFPC, Cos-phase, MWPC	SVM	0.008	3.922	1.965	19.57
Sahidullah, *et al.* (2015) [**83**]	Short-term power spectrum, short-term phase	GMM SVM	0.22 3.43	3.87 16.76	2.045 10.095	- -
Tian, *et al.* (2016) [**84**]	HD, HF, dynamic (LMS, RLMS,IF,BPD,GD,MGD)	NN	0.124	5.224	2.62	26.10
Zhang, *et al.* (2016) [**85**]	TEO, PMVDR	DNN	0.67	6.04	3.35	27.94
Patel, *et al.* (2016) [**86**]	F_0 and SoE (MFCC+CFCCIF)	GMM	0.34	1.66	1.00	6.64
Alam, *et al.* (2016) [**87**]	DFB, DMCC, DLPCC and DPSCC	DNN	1.16	3.21	2.18	12.86
Massimiliano *et al.* (2016) [**88**]	Constant Q Cepstral Coefficients (CQCCs)	GMM	0.048	0.462	0.255	1.065
Soni, *et al.* (2016) [**89**]	SBAE + MFCC	GMM	0.502	1.83	1.16	7.283
Bhavsar, *et al.* (2016) [**90**]	LP and NLP	GMM	0.25	5.00	2.632	23.71
Patel, *et. al* (2016) [**91**]	MFCC + CFCCIFS	GMM	0.354	1.49	0.922	5.7
Qian, *et. al* (2016) [**92**]	DNN-based deep features, BLSTM-based deep features	LDA SVM	0.00	2.160	1.080	10.7
Patel, *et. al* (2017) [**93**]	S-F interaction features + MFCC	GMM	0.256	4.119	2.18	19.31

Note: '-' indicates that the details are not available

and RPS feature set. This approach achieved *0.025 %* EER on vocoded speech (known attacks). However, an EER of *40 %* was observed on vocoder-independent spoof (unknown attacks).

Very recently, the use of Mel Wavelet Packet Coefficients (MWPC) was explored for SSD task [**82**]. On the development set, these features performed better than MFCC, Mel Frequency Principal Coefficients (MFPC) and Cos-phase features. However, on the evaluation set, MWPC did not perform better for unknown attacks. In this work, the use of SVM and Deep Belief Network (DBN) was used as a classifier with SVM performing better than DBN. Simultaneously in [**83**], a comparison of various features was presented and it was observed that spectral information in the high-frequency region, dynamic information in speech and detailed information related to subband characteristics of speech are considerably more useful in detecting spoofed speech. A similar study was reported very recently in [**84**], concluding that higher-dimensional, high-frequency regions and dynamic temporal information of LMS, RLMS, IF, BPD, GD, and MGD feature sets proposed in [**72**] were significant for SSD task. Very recently, in [**85**] the nonlinearity property in the speech was explored using the Teager Energy Operator (TEO) and by using Perceptual Minimum Variance Distortionless Response (PMVDR) using both GMM and DNN classifier. They observed that using large training data as required in DNNs does capture some information on the development set. However, not much improvement is observed on the evaluation set as spoofing attack can be unknown and not also present in the training data. Using the TEO and PMVDR features jointly an EER of *0.67* % was obtained for known attacks and *6.04* % for unknown attacks and with *27.94* % EER for *S10* spoof which is a unit-selection vocoder-independent spoof.

In [**87**], another approach with DNN-based classification system was presented where Delta Filterbank spectra (DFB), Delta plus double delta Mel Frequency Cepstral Coefficients (DMCC), Delta plus double delta Linear Prediction Cepstral Coefficients (DLPCC) and Product Spectrum-based Cepstral Coefficients (DPSCC) features were used. A DNN is trained on the spoofing challenge training and for each feature, posteriors and Bottleneck Features (BNF) are generated. The DFB-BNF, DMCC-BNF, DLPCC-BNF, DPSCC-BNF and DPSCC-DNN systems gave *0.013* %, *0.007* %, *0.0* %, *0.022* %, and *1.00* % EER, respectively, on the *S1-S9* spoofing attacks

and *32.28* %, *33* %, *32.69* %, *21.47* % and *12.86* % EER for *S10* spoof, respectively. Very recently [**92**], to incorporate deep learning into spoofing detection outputs from the hidden layer of various deep models are employed as *deep features* for spoofing detection. The DNN-based frame-level and Recurrent Neural Networks (RNN)-based sequence-level feature extraction framework are explored. The EER of the best deep feature system achieves nearly *0.0* % for all vocoder-based *S1 to S9*, and gets *1.1* % on *S1-S10*, which is very promising performance in ASVspoof2015 Challenge task.

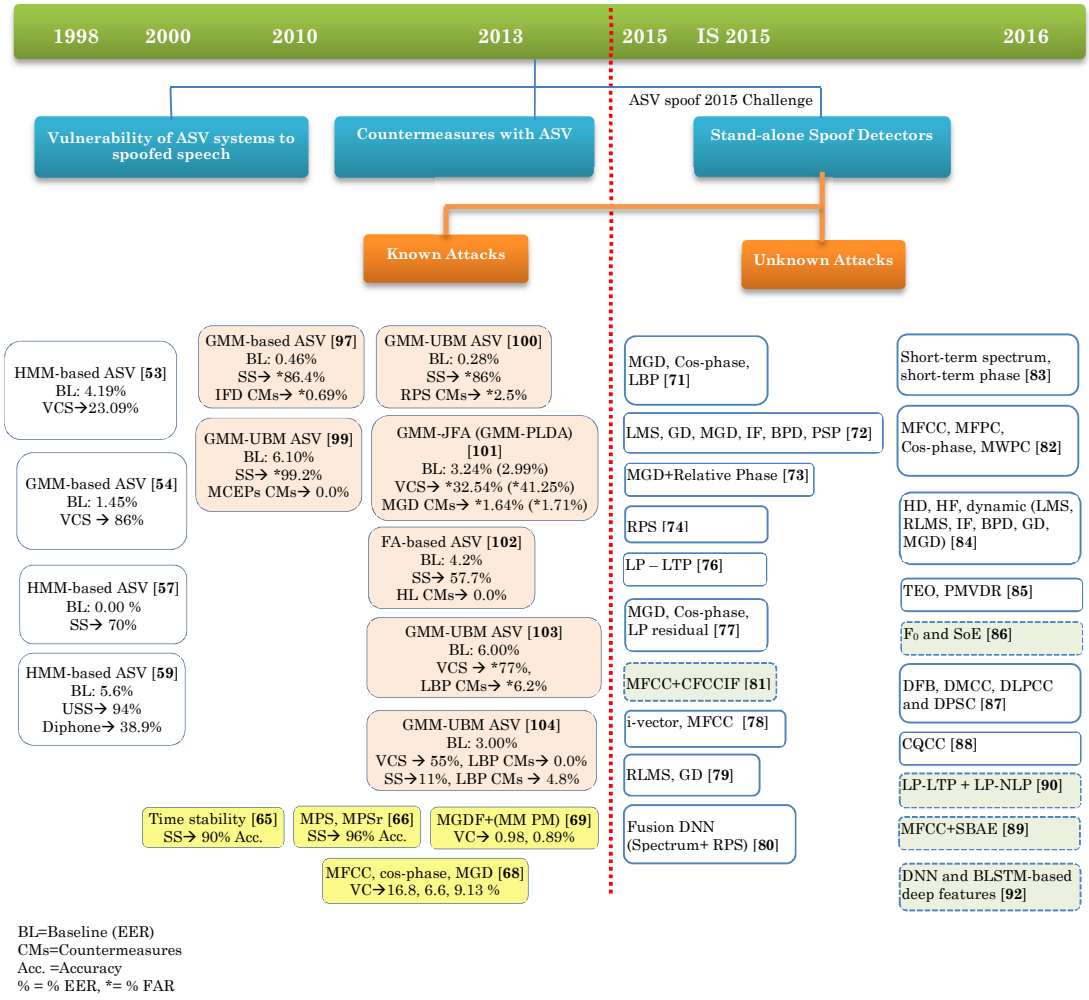


Figure 2.3: Summary of literature of SS and VCS spoof detection

Very recently, the authors in [**88**] proposed feature set based on Constant Q Cepstral Coefficients (CQCCs) which ensure a constant Q factor across the entire spectrum resulting in a higher frequency resolution at lower frequencies while providing a higher temporal resolution at higher frequencies. This reflects more closely the

human perception system. This work has reported the least EER of *1.0* % on the *S10* spoof. It is demonstrated that the performance of SSD system is more dependent on the particular features used rather than on the particular classifier.

2.4.3 Countermeasure for Signal Degradation Conditions

Amongst the several approaches used for spoofed speech detection, the results are evaluated under clean condition. To mimic the real life scenario, the research has been directed towards evaluating the performance of countermeasures under noisy environments. Table 2.3 presents a summary of the work done for spoof detection in noisy environments.

Table 2.3: A summary of the various countermeasures used for spoof detection under noisy conditions on the ASV spoof 2015 challenge database

Study	Type of Noise	SNR (dB)	Feature Sets	Classifier
Tian, *et al.* (2016) [**94**]	White, Babble, Volvo, Street, Cafe	20, 10, 0	LMS RLMS IF BPD GD, MGD	MLP
Tian, *et al.* (2016) [**95**]	White, Babble, Volvo, Street, Cafe, reverberation	20, 10, 0	LMS RLMS IF BPD GD, MGD	MLP
Hanilci, *et al.* (2016) [**96**]	White, Babble, Car	20, 10, 0	MFCC, IMFCC, SCMC, MHEC, RPS MGD, Cos-Phase	GMM, PLDA
Patel, *et al.* (2017) [**93**]	White, Babble, Car	20, 10, 0	S-F interaction features	GMM

In [**94**], a preliminary investigation of spoofing detection under additive noisy conditions with Multilayer Perceptron (MLP) classifier and features as in [**72**] is presented. It describes an initial noisy database developed by artificially adding background noises at different Signal-to-Noise Ratio (SNR) to the ASVspoof challenge 2015 database. The work shows for a model trained on clean data, the system performance degrades significantly in noisy conditions. They conclude that the systems performance differs with the type of noise. The phase-based features were found to be more noise robust than the magnitude-based features. In an extended study, reverberation noise with different reverberation time is considered [**95**]. It is observed that the reverberation noise does not affect the performance of the SSD system to a large extent. In [**96**], on the similar grounds, using seven acoustic feature sets (i.e., *4* spectral magnitude and *3* spectral phase related features), it was observed that the countermeasures break down even at relatively high SNRs and fail to generalize for spoof detection even with speech enhancement. A simple GMM back-end was found to be relatively best. Although the performance

of features depends on the type of noise, on an average, the MFCCs and Subband Spectral Centroid Magnitude Coefficients (SCMCs) performed best. The details of S-F interaction features shown in Table 2.3 are given in the next sub-Section.

2.4.4 Contributions in the Thesis in Relation to the Literature

This sub-Section discusses the contribution in the thesis relative to the literature (as shown by dotted regions in Figure 2.3). To compare the performance of the features proposed in this thesis, the results shown in Table 2.2 are reported for spoof-dependent threshold. In this case, the EER is estimated for natural *vs.* each of the spoofing algorithm and then the average EER is taken as the mean of all the EERs. However, in the rest of the thesis for the ASV spoof challenge database we use threshold independent approach to estimate the EER (as discussed in Chapter 3). This is a much realistic case and it has also been used and accepted as a performance measure in the recent ASV spoof 2017 challenge [**22**].

For the challenge, as discussed in Chapter 4 of the thesis, the Cochlear Filter Cepstral Coefficients plus Instantaneous Frequency (CFCCIF) features were proposed with *0.4* % EER for the known attacks and the relatively best EER of *2.013* % for unknown attacks [**81**]. Further improvements were observed on using CFCCIFS that gave an average EER is *0.922* % with *5.7* % for the *S10* spoof which is an improvement over the MFCC-CFCCIF system [**91**]. Thereafter, on the lines of using DNN-based approach, an Autoencoder (AE) network that embeds the subband processing of the human ear, i.e., Subband Autoencoder (SBAE)-based features is used for the SSD task [**89**]. These features after fusion with MFCC gave very less EER of *7.28* % for the *S10* spoof (details of this approach is given in Chapter 4). In [**86**], a source-based approach was presented that used F_0 and Strength of Exciation (*SoE*) features along with score-level fusion with MFCC and CFCCIF features to achieve as low as *6.6* % EER for *S10* spoof. Other source-based approch includes a modification to the existing LP-LTP approach [**76**] that uses Non-Linear Prediction (NLP) to detect spoofed speech [**90**] (details of these approaches are presented in Chapter 5). In [**93**], the use of S-F interaction based features was proposed for the spoof speech detection task. This makes use of the residual difference between the estimated glottal flow waveform and it's fitted LF-model. The residual is represented in both time and frequency domain. As shown in Table 2.2, the S-F interaction

features with MFCC obtain an EER of *0.25* % for known and *4.119* % for unknown attacks. The time-domain residual and shape features were found to be robust to signal degradation conditions and gave promising results than MFCC. The time-domain features are at much lower frequency as a result of which the degradation due to high frequency component is less. It is observed that the frequency-domain features are highly sensitive to noise and hence, not suitable for spoof detection in noisy conditions (details of this approach is given in Chapter 6).

2.5 Countermeasures in Conjunction with ASV Systems

To prevent the impostor from deceiving the ASV systems, research has been directed in developing countermeasures and jointly using it with the ASV systems to detect SS or VCS spoofed speech. In the joint ASV and spoofing detection approach, a baseline ASV system is developed with a certain % EER. Thereafter, the zero-effort impostors are replaced with the spoofed speech and the % FAR is noted at the % EER of the baseline ASV system. The countermeasures are then used with the ASV system to achieve lower % FAR than that obtained with spoofing. Table 2.4 shows the summary of ASV performance integrated with spoofing countermeasures.

Most approaches available in the literature to detect synthetic speech rely on detecting and dealing with artifacts specific to the synthesis or spoofing algorithm. Generally, countermeasures are designed based on the observation that the synthetic speech tends to have less dynamic variation in the speech parameters than those of natural speech signal. In [**97**], use of Intraframe Differences (IFD) as a discriminative feature was proposed. The reason behind this is that in the HMM-based speech synthesis system, the speech parameter sequence is generated to maximize the output probability and hence, the variation in likelihood will be less than that of the natural speech signal. Without any prior synthetic speech discrimination, the FAR was as high as *86 %* which reduced to *0.69 %* by using IFD features. This method detects well HMM-based synthetic speech generated using [**98**]. In [**99**], higher-order Mel Cepstral Coefficients (MCEPs) are used to detect HMM-based synthetic speech. The variance of *14*th to *24*th order coefficients was used as a discriminative system. The higher-order cepstral coefficients are smoothed during HMM model parameter training and synthesis process. Therefore, the higher-order coefficient of synthetic speech reveals less variance than that of the natural

Table 2.4: A summary of the various countermeasures used for SS and VCS spoof detection jointly with the ASV systems prior or post verification

Study	# of Speakers	ASV System	Spoof Type	Before Spoofing	After Spoofing		With CMs	
				% EER	%EER	%FAR	%EER	%FAR
Satoh, *et al.* (2001) [**97**]	20	GMM	HMM-SS	0.46	27.1	**86.4**	-	0.69
Chen, *et al.* (2010) [**99**]	14	GMM-UBM	HMM-SS	6.10	-	**99.2**	0.00	-
De Leon, *et al.* (2012) [**100**]	WSJ 283	GMM-UBM SVM	HMM-SS HMM-SS	0.28 0.00	- -	**86** **81**	- -	2.5 2.5
Wu, *et al.* (2012) [**101**]	NIST '06 504	GMM-JFA	GMM-VCS	3.24	7.61	17.36	-	0.00
			USS-VCS		11.58	32.54	-	1.64
		PLDA	GMM-VCS	2.99	6.77	19.29	-	0.00
			USS-VC		11.18	41.25	-	1.71
Alegre, *et al.* (2012) [**102**]	NIST '04 201	FA_m1	Artificial	4.80	64.2	-	0.00	-
		FA_m2	Artificial	4.20	57.7	-	0.00	-
Alegre, *et al.* (2013) [**103**]	NIST'06 298	GMM-UBM	VCS	6.00	-	77	-	6.2
			SS		-	**82**	-	0.8
			Artificial		-	**91**	-	0.0
Alegre, *et al.* (2013) [**104**]	NIST'06 298	GMM-UBM	VCS	3.00	22	55	-	4.10
			SS		10.4	11	-	0.00
			Artificial		7.6	4.8	-	0.00
Z. Wu, *et al.* (2015) [**105**]	NIST '06	PLDA	VCS(GMM)	0.55	-	~14.17	-	0.00
			VCS (FS)					1.42
Dhanush *et al.* (2017) [**106**]	SAS	i-vector JFA	S1-S10	-	27.68	-	5.4 4.98	-

CMs=Countermeasures, '-' indicates that the details are not available

speech signal. However, this method also uses the prior knowledge that the spoofing attack might be from the HMM-based system and thus, the HMM parameters will already be smoothed leading to less variance. In an earlier work [**67**], experiments were carried using the WSJ corpus that consists of *283* speakers and using GMM-UBM and SVM using Gaussian supervector-based ASV systems. Using a state-of-the-art HMM-based speech synthesizer, the FAR was shown to increase from *0.35 %* to *92 %* and *96 %* for the GMM-UBM and SVM systems, respectively. In this work, after the verification process, classification of SS speech is performed using the RPS feature set. The RPS feature detected natural and SS with an accuracy of *95 %* and *88 %*, respectively. In an extended work reported in [**100**], with better ASV systems and using SS detection prior to verification, it was observed that with RPS features even after spoofing the FAR of the ASV systems were as low as *2.5 %*. The use of RPS is generally due to the acoustic differences between vocoders used in generating speech and natural speech. It should be noted that the approaches considered in the detection of HMM-based SS were based on the prior knowledge of a specific HMM-

based speech synthesis system. The same countermeasures may not generalize well to other spoofing attacks generated with other algorithms.

Considering the countermeasures developed for voice conversion spoof, an initial work was carried out using the standard NIST database [**101**]. In this work, state-of-the-art ASV systems based on GMM-JFA and PLDA are considered with % EER of *3.24* and *2.99*, respectively. The authors use their previous work where they consider the fact that for both the GMM-VCS and USS-VCS, the phase information is lost in the vocoder during the synthesis step [**68**]. Hence, phase-based features would be effective in detecting the voice converted speech. In [**68**], the DCT of the MGDF, i.e., MGDF-phase is used as features. With MGD-phase features, it was observed in [**101**] that after spoofing, the % FAR of the GMM-JFA and PLDA system reduces from *17-19 %* to *0.00 %* for GMM-VCS spoof and from *31-43 %* to *1.6-1.7 %* for USS-VCS spoof, respectively.

Major work in the development of countermeasures for voice conversion speech was carried out by F. Alegre [**102**]- [**104**]. The artificial signal attack shown in Table 2.4 is a modified of the voice conversion algorithm [**102**]- [**104**]. The work in [**102**] uses a GMM-UBM system and Factor Analysis (FA) system developed using Linear Frequency Cepstral Coefficients (LFCCs) with static and delta features (m_1) and static, delta and delta-delta features (m_2). For the FA-based ASV systems developed with m_1 and m_2 features, on spoofing with VCS, the % EER increase to as large as *64.2 %* and *57.7 %*, respectively. In this work, it is observed that adding dynamic (i.e., transitional) features makes the ASV system more robust. Here, anti-spoofing measures based on higher-level features and voice quality assessment features were used which decreased the % EER to *27 %* and *0.00 %*, respectively. For GMM-UBM baseline systems, similar results are reported using countermeasures. Thereafter, in [**103**] F. Alegre *et al.* proposed a novel countermeasure using LBP that is based on the hypothesis that process involved in generating the spoofed speech might have tampered the spectro-temporal texture as present in natural speech. An LBP operator is a *3*x*3* kernel that assigns a binary code to each pixel based on the intensity of the surrounding pixels. The LBP analysis applied on *cepstrogram* (concatenated LFCC) into a textrogram [**26**]. The LBP countermeasure is based on concatenated histogram from the pixel values across each row in the textrogram. The histograms are normalized and concatenated to form a supervector and feature set

for spoof detection. With LBP features, the % FAR for ASV system dropped down to *0.8 %* from *82 %* for SS spoof and to *6.2 %* from *77 %* for VCS spoof. The LBP features were used in [**104**] with an SVM-based classifier and better ASV system with *3 %* EER as compared to *6 %* in [**103**]. After using LBP-based countermeasures, the FAR was reduced to *0.00 %* for SS spoof and *4.10 %* for VCS spoof. It should be noted that rather than being designed for a particular spoof, the LBP is a more generalized countermeasure.

Recently, VCS anti-spoofing was performed using back-end models jointly with SV in the *i*-vector space [**105**]. This work focuses on both matched and mismatched conditions (i.e., known and unknown conditions of spoofing attacks). The authors proposed the use of speaker verification jointly with the anti-spoofing in the *i*-vector space, which allows possible integration of the two without using any fusion techniques. The detection was carried out on the NIST 2006 speaker recognition evaluation set. The % FAR after using countermeasures was as low as *10* times than that before using the countermeasures. In [**107**], two approaches of integrating the ASV system and the countermeasure, i.e., cascaded and parallel are reported. The fusion of several countermeasures is considered to offer better spoof detection. The cascaded combination of ASV and countermeasures greatly reduces the FAR whereas the FRR relatively unaffected. The parallel integration of ASV and countermeasures gives better performance when subjected to spoofing attacks. However, the performance deteriorates in the absence of spoofing [**107**]. Several results are obtained on the ASV spoof challenge database separately for male and female speakers and using different ASV systems and countermeasures. Recently, a factor modeling approach has been proposed, where the spoof variability subspace and speaker variability subspace are jointly trained [**106**]. The *i*-vector and JFA methods are used. The score-level fusion of ASV system and spoof detection system is considered which gives better performance than standalone ASV systems. On the SAS database for the *S1-S10* spoof, an average of *8.41* % EER is reported by score-level fusion than the *21.85* % EER obtained by the standalone ASV system. The use of *i*-vector and JFA methods decreases the EER to *5.4* % and *4.98* %, respectively.

The studies presented in this Section indicate that there exist various counter-measures in the literature to detect SS and VCS attacks. These countermeasures are evaluted by considering the effect on % EER or % FAR before and after spoofing.

Most of the countermeasures when used with ASV systems showed improvement in performance even in presence of spoofing. Initially, most of these countermeasures were designed and tested for known attacks, however, the research has been directed to unknown attacks due to the availability of ASV spoof challenge and SAS database. The development of countermeasures in conjunction with the ASV system is the immediate application of spoof detection task and is an upcoming research area.

2.6 Research Issues in SSD Task

Based on the literature presented in this Chapter about spoofing attacks (with and without ASV systems for known and unknown attacks), various research issues or gaps in understanding of the SSD task can be bought out. The major research issues prior to the development of the SAS database and the ASV spoof challenge were the designing of the features given that the type of attack is known. Other research issues are as follows:

- Generally, vocoder-based spoofs are considered in the spoofing database, which is not the case always. With the development of phase-aware vocoders (such as AHOCODER [**108**]), the phase-based features need modification. Thus, features other than phase-based needs to be explored.
- Majority of the features such as group delay-based features are generally motivated by other speech processing applications such as speech recognition, formant estimation, epoch extraction, etc. However, in this case, the features need to be designed specifically for the SSD task.
- Either system-based features or source-based features are generally used without directly using the source-system interactions that indeed is a vital part of the human speech production mechanism.
- Generally, for spoof detection, the FAR is considered. However, the features should have lower FRR when used with ASV systems to provide better user convenience by lesser rejections of genuine trials.
- Limited work is carried to evaluate the performance of the features in terms of robustness to signal degradation and channel mismatch conditions.
- Very few studies consider the real case scenario wherein only natural speech will be available and corresponding spoofed speech needs to be generated to build models for spoofed speech.

Considering the various research issues that exist in spoof speech detection task, in this thesis, we address the issue of tackling unknown vocoder-based spoof with less % EER. With the development of features that consider the differences between natural and spoofed speech in terms of speech production mechanism, the detection of non-vocoder spoof has also been done to a great extent. Both the source-based and system-based features that explore the speech production are considered. We explore the features derived from the nonlinear interaction between the source and system to obtain better detection performance. The S-F interaction-based features gave promising results under signal degradation or noisy conditions. In addition, we conduct preliminary experiments to evaluate the performance of the countermeasures to channel mismatch conditions. The features considered in this work obtain less % FAR and % FRR, enabling its use both as a stand-alone detector and with the ASV systems.

2.7 Chapter Summary

This Chapter described details of the various spoofing attacks, motivation and literature towards choosing SS and VCS spoof detection for the security of the ASV systems. The literature shows that previous studies were generally based on known attack detection, however, with the ASV spoof challenge database, the work in unknown spoof detection has matured. The various issues with current approaches are discussed that needs to be addressed in near future. In the next Chapter, the spoofing techniques and spoof detection system along with the various databases used in this thesis, classification system and the performance measures for the evaluation of countermeasures in the SSD system is discussed.

Chapter 3.
Spoofing Techniques and Spoof Detection System

3.1 Introduction

This Chapter discusses the two main spoofing techniques used in this study, i.e., speech syntheis and voice conversion. The technical details for developing Text-to-Speech (TTS) systems and voice conversion system are discussed along with the different aspects of Synthetic Speech (SS) and Voice Converted Speech (VCS) that contributes to unnaturalness in the speech signals. The main motivation behind this is to understand and analyze the differences between natural *vs.* spoofed speech. Next, the overall arhictecture of the spoof speech detection system along with each of the individual blocks are discussed. This includes the various databases that have been used in this study, i.e., the ASV spoof 2015 challenge database, two evaluations of the Blizzard Challenge, one in English (Blizzard Challenge 2012) and another in Indian Languages (Blizzard Challenge 2014). Thereafter, as a part of the architecture, the GMM-based classification system is discussed along with Detection Error Tradeoff (DET) curve and Equal Error Rate (EER) as performance measures.

3.2 Details of Spoofing Techniques

In this Section, we present a brief overview of the two machine-generated spoofing techniques used in this study, i.e., TTS synthesis and voice conversion.

3.2.1 Text-to-Speech (TTS) Synthesis

Speech synthesis is generally referred to as TTS synthesis technique to generate intelligible and natural-sounding speech for any given input text. The two main TTS techniques include the Unit Selection Synthesis (USS) and Statistical Parametric Speech Synthesis (SPSS) approach using Hidden Markov Models (HMMs). Recent advances are also made in the field of using DNN [**109**] and RNNs [**110**] for speech

synthesis. However, the available standard databases use the USS and HMM-based techniques and hence, we discuss their brief details in the next sub-Section.

3.2.1.1 Unit-Selection Synthesis (USS)

The USS system is developed by the Festival framework as shown in Figure 3.1. A detailed documentation of the development of USS-based TTS system can be found in [**111**]. This system has the following modules:

Text Processing: The text processing module includes handling the input text or processing the text collected from several online sources. The collected text is cleaned and several non-standard symbols, words, punctuations marks, abbreviations, tags, smileys, etc. are removed or converted to their standard form.

Phonetic Analysis: The phonetic analysis refers to the generation of a sequence of speech sound units from the text. This can make use of a dictionary as in the case of English, where the mapping from orthography to pronunciation is not always straightforward. In addition, language rules, i.e., Letter-to-Sound (LTS) rules can also be used as done in Indian languages.

Prosodic Analysis: Prosodic analysis includes using intonation and duration modeling for the given text [**112**].

Speech Generation: The speech generation can be done using rule-based, concatenative or by using a statistical approach. In general, the Festival framework uses phoneme as the basic speech sound unit [**111**], which can be modified to other speech sound units as well. In cluster unit-selection, speech sound units can be clustered based on acoustic distance. Each of the speech sound units in the data is clustered into similar acoustic groups based on the information at synthesis time, i.e., phonetic context, prosodic features and other higher-level features such as the position of a word, stress, accent, etc. To cluster units, an acoustic measure is defined in [**32**], [**113**] from acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), fundamental frequency (F_0), and delta cepstrum. The acoustic distance between the two units belonging to the same class and of different frame length is defined by a weighted *Mahalanobis distance* metric [**113**]. This measure gives the mean weighted distance between units with the shorter unit interpolated to the longer unit. In this context, Classification and Regression Trees (CART) are used to represent the clusters in Festival framework. The CART decision tree for each type

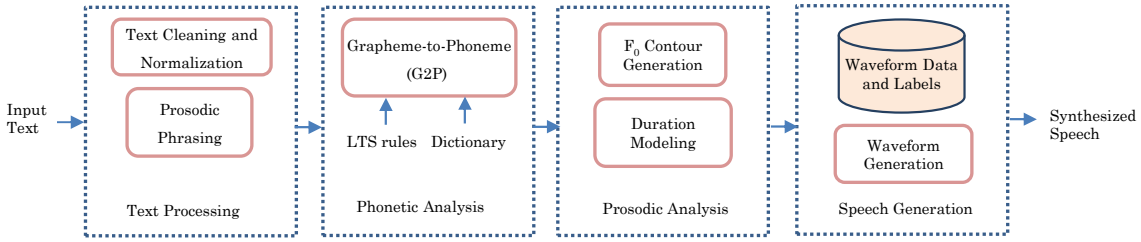


Figure 3.1: Block diagram of USS-based TTS synthesis system using the Festival framework. Adapted from [114**].**

of speech sound unit minimizes the distance of the sub-clusters at that point in a tree. The leaves of this decision tree are the list of candidate units which may be picked during concatenation, i.e., every acoustic unit in the database is associated with its symbolic context. Figure 3.2 shows an example of the decision tree for mapping context to units. The CART trees can be built based on a different feature of speech unit such as the position of speech unit in word, phrase, sentence, context of the speech unit, type of speech unit, etc. The questions are split based on the context of the current and neighboring units. Acoustic observations at the resulting leaves (i.e., context-dependent units) are the basic synthesis units for an input context. At run time, the symbolic context is extracted from the input text, the tree is traversed and once a leaf is reached, the members are extracted as the candidates.

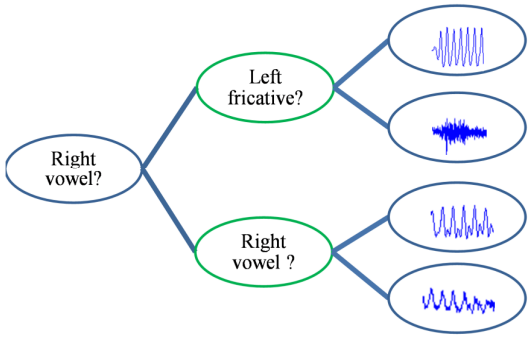


Figure 3.2: Illustration of a decision tree considering left and right context. After [114**].**

During synthesis, it is required that the best sequence of units from all the possibilities in the database is selected and concatenated. The selection comprises of two parts, namely,

- To find units in the database which best match this target unit,
- To find units which join together smoothly.

In the USS framework, to search for the best possible sequence, Hunt and Black suggested a concept of a *target cost* $T(u_i,s_t)$, i.e., how closely a database unit, u_i, is associated with the desired target unit, s_t, and a *join cost* $J(u_t,u_{t+1})$, i.e., how effectively two adjacently selected units are joined together. For each of the target unit in the database, the selection algorithm initially considers a list of candidate units, i.e., U_i. The candidate units which minimizes the sum of target cost and join cost is then selected as the final choice. The candidate speech units $\hat{U}$ are searched that minimizes the following expression [**32**]:

$$\hat{U} = \arg\min\left\{\sum_{t=1}^{N} T(u_t, s_t) + \sum_{t=1}^{N-1} J(u_t, u_{t+1})\right\}. \quad \textbf{(3.1)}$$

The waveforms of these speech sound units are then concatenated to produce synthetic speech. A Viterbi search is used to estimate the *optimal path* to obtain the lowest possible cost. The input text utterance is first parsed into a sequence of largest speech sound units (i.e., first syllable and then phone) present in the list of available units. Even with a large database, all the possible units in a language cannot be covered and only the unit present in the speech database is accessible for synthesis. Therefore, to synthesize all possible text, one instance of each sound unit must be available. However, unlike the number of phones, the numbers of larger units are not fixed. Therefore, in order to synthesize missing units, it is possible to substitute the missing unit with that of a similar unit. For example, if diphone /*d-ih*/ were missing, it could be replaced by a similar diphone of /*d-iy*/. This substitution may be perceived by the listener or may even result in recognizing a wrong word. An alternative to this is known as unit back-off where new units are made from the existing units. Instead of unit substitution, the second half part of /*t-ih*/ could be used with first part of /*d-iy*/ to create /*d-ih*/. That is, using the half-phone case [**115**].

The Festival framework uses pronunciation rules for converting the text into a *sequence* of speech sound units. After obtaining the units from the parser, the linguistic analysis module generates phonetic and related contextual features linked with each unit. If the prosodic model is built during voice building, then the target prosody is generated for the text by the prosodic analysis module. The synthesis generation stage is briefly described here. The detailed description including speech prosody modification can be found in [**114**].

3.2.1.2 Hidden Markov Model (HMM)-based Speech Synthesis System (HTS)

This HTS synthesis framework is divided into *training* part and *synthesis* part as shown in Figure 3.3. Brief descriptions of these blocks are given as follows [**116**].

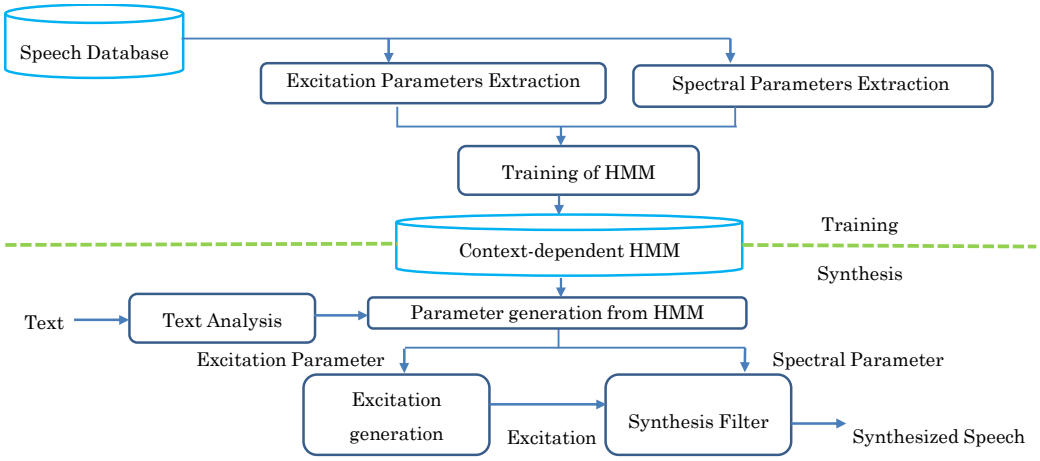


Figure 3.3: Basic block diagram of HMM-based TTS synthesis system. Adapted from [116**].**

Training part: The spectral and excitation parameters are extracted from speech database. The MFCC along with their dynamic features are generally taken as *spectral* (i.e., *vocal tract system*) parameters and log(F_0) and it's dynamic features are taken as *excitation* (i.e., *speech source*) parameters. These features are modeled by context-dependent HMMs in a unified framework [**33**].

Synthesis part: Given a test sentence which is to be synthesized, its corresponding utterance is converted to context-dependent phoneme sequence. According to the phoneme sequence, utterance HMM is constructed by concatenating context-dependent HMMs followed by determination of state duration of HMMs. Thereafter, using speech parameter generation algorithm, spectrum and excitation parameters are generated [**117**]. Finally, the speech waveform is generated using Mel Log Spectrum Approximation (MLSA) filter [**118**].

3.2.2 Voice Conversion

Voice conversion technique modifies the speech spoken by one speaker (i.e., source speaker) to give an impression that it was spoken by another speaker (i.e., the target speaker) [**38**]. Majority of the existing voice conversion systems deal with converting the spectral features of the source to match that of the target. However, prosodic

features, such as F_0 dynamics and rhythm, also contain cues of speaker identity. Similar to HMM-based speech synthesis, the voice conversion framework is also divided into training part and synthesis part as shown in Figure 3.4.

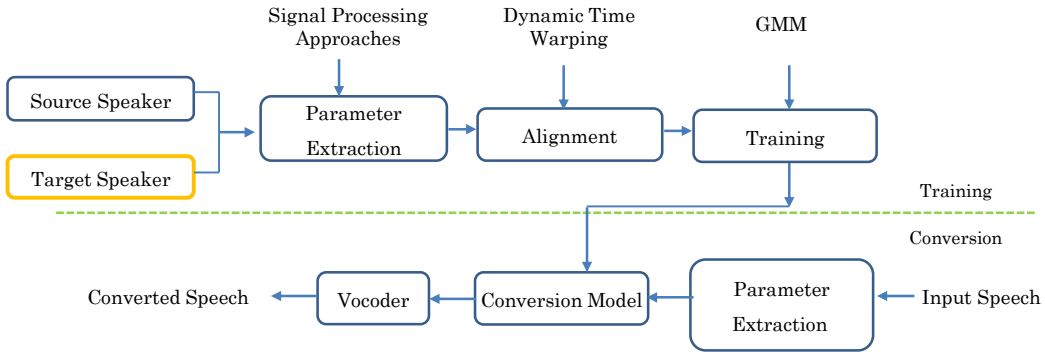


Figure 3.4: Basic block diagram of the voice conversion framework. Adapted from [119**].**

Training part: In the training part, the system is given a parallel speech (text-dependent) or non-parallel speech utterances (text-independent) from the source and target speaker. Popular speech representations are based on the source-system model where the glottal airflow acts as an excitation signal in the form of the pulse train (for voiced sounds, the glottis closes quasi-periodically) and noisy signal (for unvoiced sounds). The voiced excitation is characterized by an F_0 that is determined by the oscillation frequency of the vocal folds. The vocal tract acts a resonator cavity (having its resonances at formant frequencies) and does the job of spectral coloring of the excitation source. For the voice conversion task, features such as Linear Prediction Coefficients (LPC), Line Spectral Pair (LSP), Mel Cepstral Coefficients (MCC), Speech Transformation and Representation using Adaptive Interpolation weiGHted specTrum (STRAIGHT) analysis/synthesis framework, etc. are used [**119**]. In this phase, acoustic features are extracted both from source and target speakers. The acoustic representation of the utterances is usually at the frame-level. The representation are said to be *local* if they refer to a feature of a single frame (e.g., instantaneous pitch, energy, and spectral envelope) and *global* if they relate to an entire utterance or model of the speaker (means and standard deviations of F_0 or energy measurements, or estimates of the glottal pulse and vocal tract) [**120**]. Using the aligned features, a mapping function is learned by training the model.

Synthesis part: In this phase, the extracted speech parameters of source speaker are mapped to get target-like speech parameters. This phase is controlled by a

conversion rule obtained by a training phase. The modified parameters are used to reconstruct the new speech which shall have the characteristics of the target voice as well. Here, we discussed very briefly the voice conversion framework. The detailed description of the same using various algorithms will be discussed in Section 3.4.1.

3.2.3 Evaluation of Speech Quality

To maintain *naturalness* and *intelligibility* is the main objective of any speech synthesis or voice conversion technique. The voice quality evaluation of synthesized or converted speech is mainly done by subjective and objective measures. Next, we briefly discuss these measures to show that there are various approaches to study how the degradation in speech synthesis and voice converted speech are evaluated. This will possibly assist in using this knowledge for the design of countermeasures.

3.2.3.1 Subjective Evaluations

The subjective evaluation includes listening tests samples (by at least *20-25* subjects) where the synthetic or converted speech is scored by the listeners (subjects) to judge the naturalness, intelligibility or speaker similarity.

Mean Opinion Score (MOS): The MOS score evaluation is used both in TTS and voice conversion. In the MOS test, subjects rate the synthesized speech on a scale of *1* to *5* to evaluate naturalness. As suggested in ITU-T P.85 [**121**], score *1* stands for bad quality and *5* means excellent quality. The mean of all opinions from all the subjects is considered as the *score* for a given TTS system. Other factors that can be considered for evaluation are intelligibility, pleasantness, listening effort, articulation, pronunciation, speaking rate, overall impression, etc. [**121**].

Degradation MOS (DMOS): In DMOS the voice quality is evaluated with reference to the original natural speech signal. The subjective score obtained for synthesized speech is normalized to that of the natural speech signal. This subjective MOS is called as degradation MOS (DMOS). The listeners compare the synthetic speech quality with the reference natural speech to assess the degree of degradation. Hence, DMOS measures voice quality relative to the natural speech signal.

Semantically Unpredictable Sentences (SUS): It is used to evaluate the intelligibility of TTS system at word-level and sentence-level. The SUS test sentences are syntactically normal and semantically abnormal (with no semantic

dependencies between words to contribute what is emphasized at word-level). For example, "He is in a notebook with shoes". The listeners write whatever is heard in the sentence. In SUS test, the intelligibility of words without the sentence context is tested and this test avoids listeners to write an expected word. A general algorithm that discusses the process of SUS tests is given in [**122**].

Preference Tests: In these tests, subjects decide among systems depending on the questions asked. For example, "which system do you prefer in naturalness?" or "which system do you prefer in intelligibility?" Examples of such tests include *AB* tests and *ABX* tests. In *AB* test, during the listening task, a pair of stimuli of two different systems *A* and *B* is presented and subjects are asked to give preference. In *ABX* test, one reference stimuli *X* (corresponding to natural speech signal) and a pair of stimuli *A* and *B* are presented and subjects are asked to judge which of *A* or *B* is closest to *X* in terms of naturalness, etc.

In [**123**], instead of evaluating the quality of TTS systems, the human listening tests were carried out for the spoof detection task (detailed discussion of this is presented in Section 7.2.4). Several issues exist in subjective tests, such as the time and cost involved in running listening tests, availability of volunteers for listening tests (there may be a need to hire paid subjects). For meaningful statistics, more subjects are required which is difficult, time-consuming and also very costly. An important issue is that these tests are not always reproducible, i.e., the results may not be similar when the same test is repeated with the same listener. This is because, there exist varying cognitive factors such as listener's attention, personal feelings, not knowing the objective of the test, listening environment, not listening to the entire sentence, listeners' mood, etc. Current TTS and voice conversion systems are flexible in generating voices, therefore, every time a new voice is generated; the entire listening tests needs to be repeated again. The listening tests may not contribute much about technical details like missing attributes and lacking features in the speech signal and hence, the use of objective measures is also considered.

3.2.3.2 Objective Evaluations

Objective tests are used for simplicity and cost effectiveness. They have more shortcomings than subjective measures. In fact, they correlate less with subjective measures. Objective measures offer a measurement of voice quality by using

relevant speech signal processing algorithms. The dictionary meaning of *objective* is "not influenced by personal feelings or opinions" or "based on facts". Objective evaluations, unlike individual personal opinion, give a consistent evaluation. It also does not require time and cost for evaluation. However, there is no universally accepted objective measure for evaluation of naturalness and intelligibility. Generally, distance-based measures are used both in TTS and voice conversion.

Distance-based Measures: The Mel Cepstral Distortion (MCD) is generally used for speech quality assessment with respect to the natural speech signal [**124**]. The extracted speech features of reference and test utterances with different duration are aligned by Dynamic Time Warping (DTW) at frame-level [**125**]. The MCD between the Mel-cepstra of synthetic or voice converted speech and the natural reference speech for (N-dimension of the coefficients) is a Euclidean distance given by [**124**]:

$$MCD(k) = \sqrt{\sum_{i=1}^{N} \left[MC_x(i,k) - MC_y(i,k) \right]^2}, \quad \textbf{(3.2)}$$

where $MC_x(i,k)$ and $MC_y(i,k)$ are the i^{th} coefficients of natural and test speech, respectively. There exist other likelihood measures which include using HMM models instead of TTS systems. However, these are not found to work well and are known to be of limited use [**126**]. It is to be noted that, the distance estimated between machine-generated speech and natural speech is not the *perceptual distance* that we perceive as listeners.

Measures of Speech Prosody: Prosodic features are difficult to incorporate into any system and hence, prosodic-based objective measures would be highly effective. Prosody features include F_0 and its statistics, durational features, etc. For HTS-based systems, the quality of F_0 is measured by Root Mean Square Error (RMSE) of $\log(F_0)$ generated by HTS system and natural speech signal. This measure checks variations that govern naturalness of speech. Other prosodic features include Peakedness Ratio (number of segments having pitch changes with reference to threshold), Drop Ratio (relative number of declining F_0 contributing positively to perceived naturalness) and Variability Ratio (mean derivative of F_0 per voiced segments gives temporal variations in F_0 over time) [**126**], [**127**].

Natural speech has different *acoustic* and *prosodic* properties than synthetic or converted speech and hence, using objective measures based on these properties are

found to be useful. Thus, countermeasures for spoof detection can be proposed on these lines to identify the differences that contributed to the loss of naturalness and intelligibility in a synthesized speech. It is not clear how to define naturalness of speech signal, so both source and system features can be explored for distinguishing the naturalness of machine-generated speech from that of the natural speech signal.

3.3 Unnaturalness in TTS and Voice Converted Speech

It is known that the both TTS and voice conversion techniques lack naturalness. There are several factors that may contribute to the unnaturalness in machine-generated speech, a few of them are discussed in the next sub-Section.

3.3.1 Unnaturalness in USS-based TTS Synthesis System

Generally, USS-based TTS systems are known to be highly natural due to the direct concatenation of natural speech sound units. However, at times, the USS-based speech sound highly unintelligible due to several issues as mentioned below:

- **Labelling of speech sound units:** Labeling is a crucial step in USS voice building. Manual labeling requires a huge amount of time and efforts. In addition, manual labeling is very subjective. Automated labeling tools are available, however, not accurate enough. It is observed in [**128**], that fricative, trills and nasal sounds are highly prone to labeling errors. That is, the high energy of fricative sounds and the transient-like energy of trills sounds tend to give spurious or miss boundaries during speech segmentation task. Hence, the labeling will be inaccurate and this will severely affect the quality (in terms of overlapping sounds) of the unit-selection TTS voice.
- **The discontinuity at the joints:** Due to the concatenation of speech sound units (such as phoneme, diphone, syllable, etc.), the synthesized speech may have glitches due to abrupt joints. Although the speech sound units are selected as per minimum cost criteria, the joint needs to be smoothed in order to avoid glitches while hearing.
- **Linear phase mismatches at the joints**: As the concatenated speech sound units are recorded at various sessions, linear phase mismatches (both due to the excitation source and vocal tract system) may occur in USS voices which may be perceived during listening [**129**].

- **Lack of text-dependent speech prosody:** In the Festival framework, various parameter tuning can be carried out for the articulatory position, etc. which might affect the voice quality. The prosody in the text is also an important input parameter which if not interpreted correctly will affect the units that are chosen for joining the speech sound units.

3.3.2 Unnaturalness in HMM-based TTS Synthesis System

The biggest known drawback of HTS-based speech is that it is of *buzzy* and *muffled* quality. The various factors that may cause quality degradation are as follows [**130**]:

- **Vocoder quality (parameterization and excitation):** Previously, the vocoder in HTS was Mel cepstral vocoder with pulse or noise excitation which is way too simple than the actual speech production mechanism and hence, the synthesized speech sounds buzzy. However, with the development of HNM model, mixed excitation source and use of spectral envelope and STRAIGHT spectra, much improvement in the quality is observed.
- **Modeling accuracy:** It has been observed that there are inconsistencies in the training and synthesis procedures, i.e., the relation between the static and dynamic features are ignored in the training stage, while they are considered in the synthesis stage. In addition, with the development of Hidden Semi-Markov Model (HSMM), the inconsistency in the state duration modeling is being eliminated.
- **Over-smoothing:** The HTS synthesis is prone to over-smoothing of spectral parameter trajectories. The Maximum Likelihood (ML) parameter generation provides smooth trajectories and thus, the utterance-level variance of each parameter trajectory is significantly reduced compared to the original recordings, resulting in muffled speech. To eliminate this, several approaches like post-filtering to the enhance formant contours are used. In addition, the use of Global Variance (GV) is explored.

3.3.3 Unnaturalness in Voice Conversion System

For voice conversion, successful identity conversion is important. However, for spoofing, we are interested in naturalness which is at times deteriorated due to following reasons [**131**]:

- **Overfitting:** In GMM-based voice conversion, the overfitting can be due to training phase or when the mapping function (to convert source speaker's voice to that of target speaker) is estimated. There have been several approaches such as GMM with diagonal covariance matrices and Partial Least Squares (PLS) for regression estimation.
- **Over-smoothing:** The over-smoothing problem can occur both in time and in the frequency-domain. In the time-domain, the converted feature trajectory is much smoothened than the original target feature trajectory. Even in voice conversion, GV can be used to compensate for the reduced variance of the converted speech. In the frequency-domain, over-smoothing causes loss in finer details of the spectrum and broadening of the formants (i.e., larger *-3dB* bandwidth) and can be eliminated by post-filtering or combining the frequency warped source spectrum with the GMM-based converted spectrum to improve the quality of speech [**131**].
- **Time-independent mapping:** The GMM-based method converts each frame individually and hence, losses the temporal correlation between consecutive frames causing discontinuities in feature trajectories and resulting in degraded speech quality. Approaches like ML estimation of the spectral parameter trajectory are proposed where the static source and target feature vectors are extended with first-order deltas and a joint-density GMM is estimated and while synthesis both converted mean and covariance matrices are used to generate the target trajectory.

Amongst all the factors considered here for unnaturalness in TTS and voice conversion techniques, a basic reason is that it is not possible to exactly mimic the speech production mechanism while synthesizing speech. The use of signal processing techniques to generate excitation and spectral parameters will leave some artifacts in the spoofed speech, which are again different for different spoofing algorithms. As discussed in Section 2.5, the use of IFD feature [**97**] and features based on variance of higher-order MCEPs [**99**] explore the fact that over-smoothing is an artifact of the synthetic speech and hence, can be used for the SSD task. Other approaches such as using linear prediction based techniques, explore the idea that spoofed speech is easily predicted (if generated by a simple linear acoustic model) or difficult to predict (due to discontinuity at joints of natural speech units in USS-

based speech) [**76**]. This analysis illustrates the need to identify a common artifact in the machine-generated speech which will aid in SSD task.

3.4 Architecture of the Spoof Speech Detection System

The general SSD system can be divided into database, feature extraction, classification and decision making as shown in Figure 3.5. Next, we describe each of the standard databases that are used in this study for the spoof detection task. The details of the databases and the approach used to develop the spoofing material are discussed in detail. Thereafter, the approach used for classification system and the methodology of decision making is discussed in the next Sub-Sections.

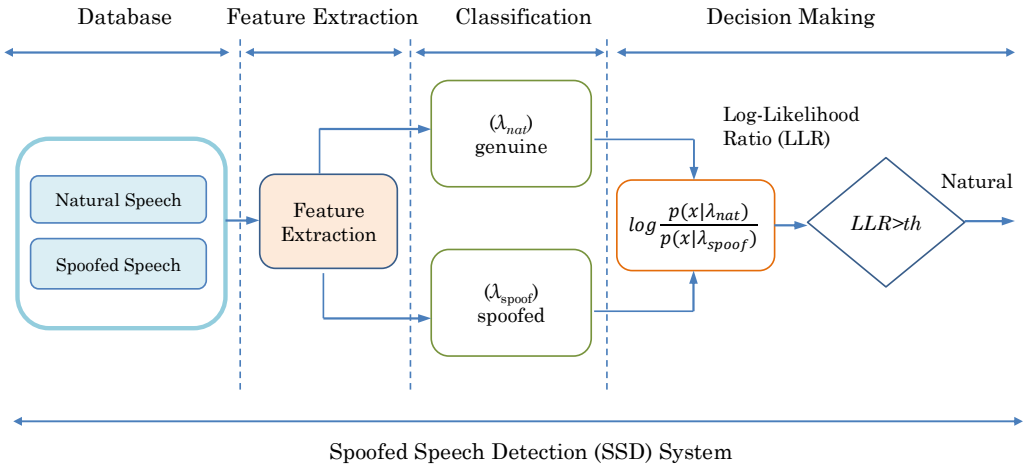


Figure 3.5: General architecture of spoof detection system used in this thesis.

3.4.1 Details of Databases

In this thesis, the ASV spoof 2015 challenge and Blizzard Challenge datasets are used. The ASV spoof 2015 challenge database consists of both known and unknown attacks [**20**]. The Blizzard Challenge 2012 database is in English language [**24**] and Blizzard Challenge 2014 database consists of Gujarati and Hindi language [**25**]. This dataset is used to evaluate the robustness of the features to channel mismatch case and observe the case of language dependency as well.

3.4.1.1 ASV Spoof 2015 Challenge Database

This dataset was provided for the ASV spoof 2015 challenge held at INTERSPEECH 2015 [**20**]. The ASV spoof challenge dataset is the subset of the Spoofing and Anti-

Spoofing (SAS) corpus which is the first such generalized and statistically meaningful dataset containing various spoofing attacks [**19**]. The database includes both speech synthesis and voice conversion spoofing attacks, which are the most accessible and highly effective spoofing approaches [**26**]. In the challenge database, *3* out of *10* spoofing algorithms were based on synthesis techniques and remaining *7* using voice conversion techniques. The speech synthesis spoofing set was developed using the SPSS and unit-selection synthesis approach. The voice conversion spoofing sets were created using one publicly available open source toolkit and *6* state-of-the-art voice conversion techniques. The dataset is divided into three sets, i.e., training, development and evaluation set. For the challenge, the task was to create a detector that could detect both known and unknown attacks. Therefore, the training and development set consists of *5* spoofing algorithms from *S1-S5* (i.e., known attacks) and the evaluation set has *10* spoofing algorithms from *S1-S10* (i.e., both known and unknown attacks). There are in total *106* speakers in the database and the speakers in each set are exclusive. The details of the number of speakers and the number of utterances in the ASV spoof challenge database are shown in Table 3.1.

Table 3.1: Summary of utterances used in training, development and evaluation sets of the ASVspoof 2015 challenge database [20**]**

Dataset	Hours	No. of Speakers		No. of Utterances	
		Male	Female	Genuine	Spoofed
Training (S1-S5)	15	10	15	3750	12625
Development (S1-S5)	48	15	20	3497	49875
Evaluation (S1-S10)	170	20	26	9404	184000

The SS spoofs were generated from two sets of data, i.e., *Part A* and *Part B* consisting of *20* and *40* utterances to train the systems, respectively. In all the spoofing (S) algorithms, unless specified, STRAIGHT was used during analysis to extract *24-D* MCCs, *25* Band Aperiodicities (BAPs), and F_0 features [**132**]. In all voice conversion techniques, unless specified, F_0 was converted by global linear transformation (simple mean-variance normalization). The brief details of the various algorithms are provided below [**20**].

S1 (VCS): It is a simplified Frame Selection (FS)-based voice conversion algorithm, in which the converted speech is generated by selecting target speech frames. It is a simplified version of exemplar-based unit-selection [**46**], using a single frame as an exemplar and without a concatenation (join) cost. The MFCC vectors are mapped

from a source speaker to the target speaker. These mapped MFCC vectors are then used to select actual frames from the speech database of the target speaker. From the selected frames, LPCs are extracted and the speech is synthesized by analysis-synthesis framework.

S2 (VCS): It is a simple voice conversion technique. The excitation F_0 was converted by a global linear transformation (simple mean-variance normalization). The BAPs were copied, without undergoing any conversion. The first coefficient of the source speaker's MCC (c1) was converted by a linear transformation [**133**]. This is the simplest voice conversion method since it only changes the overall slope of the spectral envelope and not any other speaker-specific features.

S3 (SS): This is an HTS system based on the SPSS approach as in [**35**] and with speaker adaptation framework as described in [**36**]. During speech analysis and average voice training phase, STRAIGHT vocoder was used to extract *60-D* Bark Cepstral coefficients, log (F_0) and *25-D* BAPs [**132**]. The HSMMs are trained on a large multi-speaker voice bank corpus to simultaneously model acoustic features and duration. For adaptation, speech data from *Part A* was used. To synthesize speech, acoustic feature parameters are generated from adapted HSMMs using a parameter generation algorithm that uses GV. The excitation signal is generated using mixed excitation and Pitch-Synchronous Overlap and Add (PSOLA) and STRAIGHT vocoder was used to create the final synthetic speech waveform.

S4 (SS): This system is the same as *S3*, except that speech data from *Part B* was used to train the system.

S5 (VCS): This voice conversion technique is implemented with publicly-available open-source Festvox system [**134**]. The algorithm uses a joint density GMM with ML parameter estimation [**135**]. The *Part A* set of parallel training data is used and settings of the toolkit are set to default with *32*-Gaussian components. This uses a MLSA vocoder for speech generation [**133**].

S6 (VCS): This voice conversion technique is similar to *S5* with some enhancements. The algorithm uses a joint density GMM with ML parameter generation using GV [**135**]. The *Part A* set of parallel training data is used considering MCC feature set. As in *S5* generation approach, *32* components GMM were chosen.

S7 (VCS): This spoof is similar to *S6* spoof. However, it uses LSP rather than MCC for spectrum representation.

S8 (VCS): This is a Tensor-based arbitrary Voice Conversion (TVC) system [**136**]. To construct the speaker space, the Japanese dataset was used and only the MCC were converted, without altering other features.

S9 (VCS): This system uses Kernel Partial Least Square (KPLS) regression [**137**], trained on the *Part A* data with *300* reference vectors and Gaussian kernels were used to derive kernel features. In addition, *50* latent components were used in the PLS model. Dynamic kernel features were not included, for simplicity.

S10 (SS): It is a vocoder-independent USS-based algorithm implemented using the open source Modular Architecture for Research on speech sYnthesis (MARY) Text-To-Speech (MARY TTS) system [**138**] that uses Festival framework [**111**]. As described earlier in Section 3.2.1.1, the USS framework consists of text pre-processing, linguistic analysis and annotation (i.e., part of speech (PoS) tagging, G2P conversion, etc.) and cluster unit-selection with diphone synthesis.

Thus, to summarize, the *S3*, *S4,* and *S10* are SS spoof and remaining are VCS spoofs. The *S5* VCS spoof uses MLSA filter [**133**] and other spoofs use STRAIGHT vocoder [**132**] (except *S10*). The *S10* spoof is vocoder-independent and does not use any vocoder for speech synthesis. For speech synthesis, the development set consisted of vocoder-dependent spoof while during testing, the vocoder-independent spoof was considered. This was not the case for voice conversion attacks. However, the spoofs in the training were only based on MCC and then tested for LSP-based conversion technique. Hence, *S7* and *S10* would possibly be difficult to detect than other spoofs in the evaluation set.

3.4.1.2 Blizzard Challenge Database

To compare the effectiveness of various research techniques in building corpus-based speech synthesizers on the same data, the Blizzard Challenge has been planned annually since the year 2005. The basic challenge is to build a TTS system and submit the given set of test sentences for evaluation. The submitted test sentences are evaluated through listening tests carried on the natural utterances used as a reference and the corresponding synthesized versions [**139**]. The approaches used in Blizzard challenge are diverse and includes, SPSS, USS and even a hybrid model of

the two. Thus, evaluating the performance of the countermeasures on Blizzard challenge dataset will test the robustness of the features on completely unknown algorithms and for speakers which are definitely not present in the training set. In addition, the Blizzard database has completely different recording conditions than that of the ASV spoof challenge database, and hence, it will aid to study the robustness of the features to channel mismatch case. An overview of the Blizzard Challenge and the technologies used for TTS is presented in [**140**]. We consider in this study the latest evaluation of only English language at Blizzard Challenge and the only Indian languages version of the Blizzard series. The wave files for various submissions at the Blizzard Challenge can be downloaded from [**141**].

3.4.1.2.1 Blizzard Challenge 2012 Dataset

For the Blizzard Challenge 2012, a single-speaker corpus was used, created from audiobook recordings on the Librivox website [**24**]. A Festival-based unit-selection system *B* was used as a benchmark. There were *9* teams that participated in the challenge. Therefore, we have *11* systems, i.e., from *A* to *K* under consideration. In particular, system *A* contains natural speech signals whereas systems *B*, *G*, *F* and *I* were built using unit-selection method. Systems *E*, *H*, *K* were built using statistical methods. Systems *C* and *D* were built using hybrid and *J* was built using the diphone-based method. Each system has two categories, namely, *60* paragraphs and *100* sentences. In this work, we use *100* read sentences from each system.

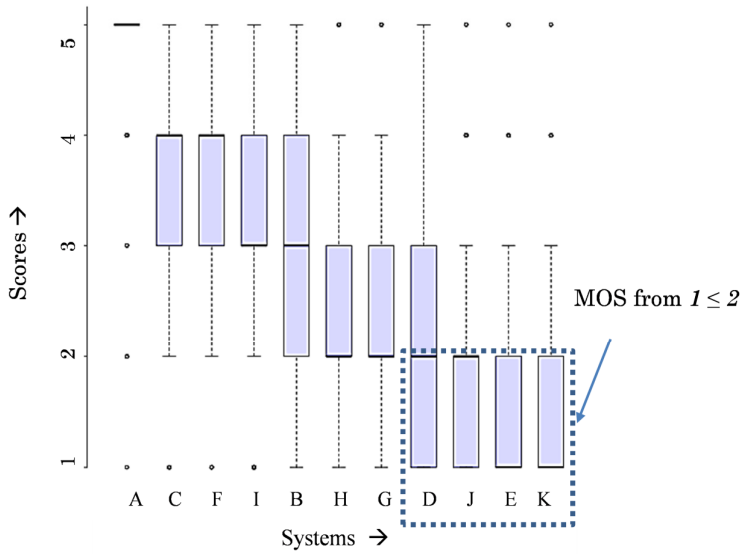


Figure 3.6: The MOS of various systems at the Blizzard Challenge 2012. Adapted from [24].

The naturalness of the systems can be known from the MOS obtained from the listeners. Figure 3.6 shows the MOS scores of various systems at the Blizzard Challenge 2012. The natural system *A* has an MOS score of *5*. It is observed that on an average, the USS systems are more natural sounding than HMM-based speech synthesis systems. As shown in Figure 3.6, systems *D*, *J*, *E* and *K* have MOS $1 \leq 2$.

3.4.1.3 Development of Gujarati Database

The Gujarati speech data is collected as a part of sponsored consortium project by Department of Electronics and Information Technology (DeitY), New Delhi, India, namely, "Development of Text-to-Speech Synthesis in Indian Languages-Phase II". The main use of TTS systems was for the visually challenged and for those having the cerebral palsy disorder. The steps in the development of TTS for the Gujarati language are discussed next.

3.4.1.3.1 Text Corpus Collection

About *200000*+ words were collected from sources such as newspaper articles, magazines, stories, essays, etc. comprising of *5651* unique syllables. To reduce the text data and still maintain high syllable coverage, text optimization was carried out to cover as many syllables as possible with a minimum number of repetitions.

- The text is divided into lines containing around *10* words.
- For each line, a score is calculated based upon the number of "*new syllables*" in the line and the "*frequency*" of those syllables. The syllabification script is used for this purpose. A "*new syllable*" means that the syllable has not already been selected by the optimization process in the optimized text.
- After each of the iteration, top *300* lines were appended to a file which would contain optimized text.
- Above process is repeated iteratively till all the lines are covered in the optimized text.

The optimized text has the highest scoring lines at the top and lowest scoring lines at the bottom. Figure 3.7 shows the relation between number of *unique* syllables (i.e., without repetition) and the number of lines. The number of syllables saturates after some lines, and this helps in deciding the text data and the number of hours of actual speech that can be used for recording based on the syllable coverage required.

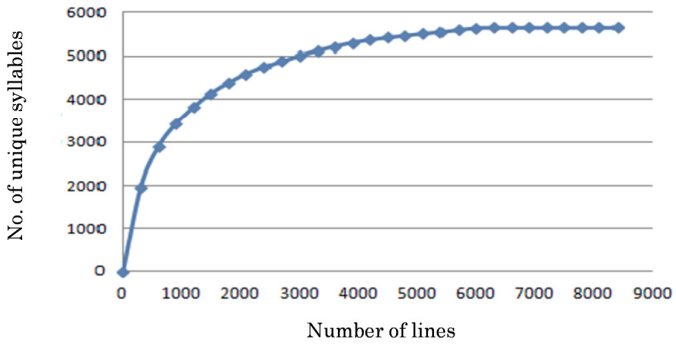


Figure 3.7: Accumulation of unique syllables for Gujarati language for the optimized corpus.

3.4.1.3.2 Voice Artist Selection and Recording

The task of voice artist selection was carried by contacting various radio stations in Ahmedabad, Gujarat State, India. The test speech samples were collected from *10* female and *5* male native voice over artists. The candidates have been explained the purpose of the recordings and the use of voices after TTS system development. As the TTS synthesis systems were meant for visually challenged and for those with the cerebral palsy disorder, the voices after synthesis should be *soothing* and should not cause long-term irritation or discomfort upon hearing. Therefore, the speakers have been chosen appropriately after performing enough signal processing experiments on the test speech samples from voice over artists. Thus, the voice over artists for speech recording were selected based on MOS evaluations obtained on the *pitch*, *tempo* and *shrillness* factor. The shrillness factor is evaluated by a visually challenged subject in terms of the Softness (SF) of the voice.

The test voice samples were altered in terms of *pitch* and *tempo*. The speech waveform after pitch and tempo changes were saved separately and played for MOS evaluation. The score was calculated based on the average of the pitch and tempo changes. Here, pitch change (by *15* %) is less as compared to tempo change (by *50* %) based on the observation from a visually challenged subject using NonVisual Desktop Access (NVDA) screen reader that they use fewer pitch changes while listening (especially decrease in pitch). It is the slow and high speed that matters before and after getting used to the TTS system voice, respectively. After considering pitch and tempo changes, the priority was considered. In the case of priority, a particular factor was given *50* % priority and the remaining factors were assigned *25* % priority each. The softness was evaluated to know the comfort during the long-

term hearing. The MOS score for the above parameters was given by listening test at a scale of *1* (i.e., very poor) to *5* (i.e., excellent). Here, MOS is taken from *11* subjects (*9+2=11*, *9* male and *2* female who are university graduates without any background in speech signal processing area). For the MOS evaluation in Table 3.2, the age of the female speakers varied from *25* to *40* years. It can be seen that *Speaker 2* (with an age of *35*) has outperformed all the other speakers in all the tests. The *Speaker 5 and Speaker 9* have better MOS for PC. However, the difference is not significant and can be neglected considering the better performance in rest of the factors (especially in softness). Hence, *Speaker 2* was selected for speech data recording.

Table 3.2: Average MOS for female artist selection from *11* subjects. Adapted from [128**]**

Female Artist	PC (+15%)	TC (+50%)	TC (-50%)	S	PP (+50%)	TP (+50%)	TP (-50%)	SF
1	3.72	2.72	1.81	2.75	2.99	2.74	2.52	**4.00**
2	**3.75**	**2.88**	**2.78**	**3.14**	**3.29**	**3.07**	**3.05**	**4.05**
3	3.72	2.27	1.54	2.51	2.81	2.45	2.27	2.25
4	3.27	2.36	1.72	2.45	2.65	2.43	2.27	3.00
5	***4.00***	2.45	1.8	2.75	3.06	2.68	2.51	2.50
6	2.45	2.00	1.27	1.91	2.04	1.93	1.75	1.75
7	3.36	2.18	1.27	2.27	2.54	2.25	2.02	2.00
8	3.63	2.36	1.81	2.60	2.85	2.54	2.40	2.85
9	***3.9***	2.45	2.00	2.78	3.06	2.70	2.59	3.65
10	3.75	2.63	2.62	3.00	3.18	2.91	2.91	3.20

*PC=Pitch Change, TC=Tempo Change, S=Score, PP=Pitch Priority, TP=Tempo Priority, SF=Softness.

Table 3.3: Average MOS for male artist selection from *11* subjects. Adapted from [128**]**

Female Artist	PC (+15%)	TC (+50%)	TC (-50%)	S	PP (+50%)	TP (+50%)	TP (-50%)	SF
1	**3.82**	**3.6**	**2.35**	**3.26**	**3.40**	**3.34**	**3.03**	**2.95**
2	4.00	3.41	2.1	3.17	3.38	3.23	2.90	2.9
3	3.33	3.08	1.7	2.70	2.86	2.79	2.45	2.1
4	1.7	3.80	2.4	2.63	2.40	2.90	2.56	3.0
5	***4.05***	*3.75*	*2.3*	*3.37*	*3.54*	*3.46*	*3.1*	*2.85*

*PC=Pitch Change, TC=Tempo Change, S=Score, PP=Pitch Priority, TP=Tempo Priority, SF=Softness

The analysis done on the male test speakers and the MOS evaluated are shown in Table 3.3. The male *Speaker 5* has the best MOS in tempo and pitch changes as well as the pitch and tempo priorities. However, taking into consideration the shrillness factor, the MOS (given by a visually challenged subject) was less. This is because the voice was *not* soothing and not appropriate for long-term hearing. In addition, considering the age factor, the *Speaker 5* had an age around *50* years as compared to the remaining speakers who were between *25-35* years of age. Therefore, *Speaker 1* was selected for recording. Each female session lasted for approximately *3.5* hours and the male session lasted for *2.5* hours. The wavefile of each session was cut and saved one-by-one according to the text and the wave files (with same name tags). The

cutting of wavefiles was done at the DA-IICT, Speech Research Lab. There was enough silence between the lines which aided the splitting task. The raw files were recorded at *48 kHz* and downsampled at *16 kHz* by using *chmod* command in Linux, the channel was *mono* with *16*-bits resolution.

3.4.1.3.3 Unit-Selection Synthesis (USS) in Gujarati

For USS system building, it is necessary to decide the basic speech sound unit on which the system is to be built. For Indian languages, syllables are best suited as the basic unit because Indian languages are syllable centered or syllable-timed [**142**]. A syllable is a unit having a vowel at the nucleus surrounded by none or more consonants. It is typically of the form *C*VC**, where *C* is a consonant, *V* is a vowel and *C** indicates none or more consonant present. The syllable consists of the onset, rime and coda. The *onset* and *coda* can consist of consonants while the *rime* consists of the vowel. Once the speech sound unit is defined for a language, the Letter-to-Sound (LTS) rules need to be identified prior to speech data labeling. The LTS rules indicate how the written text has to be spoken. The LTS rules for Gujarati are very similar to LTS rules of Hindi [**142**] and are discussed in [**143**]. For the both Hindi and Gujarati language, the written and spoken form has close correspondence. However, *inherent vowel* (i.e., short /a/) associated with each consonant is not always spoken depending on the context. This is referred to as *schwa deletion* or Inherent Vowel Suppression (IVS). For example, the word *pala* (meaning 'moment' in English) is mapped to sounds */p/ /a/ /l/*, ignoring the vowel associated with */l/*. The rules to determine IVS of a consonant character are derived from Hindi (an Indian language). The details of these rules in Gujarati are provided in [**142**]- [**143**].

Based on LTS rules, the syllabification rules are required to break the words into syllables. By analyzing several words, it is observed that the syllabification rules of Gujarati are similar to syllabification rules of Hindi [**142**]. Once the syllables are identified using the LTS and syllabification rules, the text and corresponding particular part of speech segment needs to be aligned. This is done by DONLabel labeling tool (based on minimum-phase Group Delay (GD) segmentation method) developed by IIT Madras for Indian languages [**144**]. In the DONLabel tool, the Window Scale Factor (WSF) needs to be set for each utterance such that the syllables get distributed to complete corresponding speech utterance. Thereafter,

manual adjustments are done for accurate labeling. Figure 3.8 shows the labeling done by DONLabel tool after manual adjustments. Although, we use the DONLabel tool for labeling the speech data followed by manual corrections, there are approaches proposed whereby automatic segmentation can be done to reduce the manual efforts. It has been shown that by using approaches like Gaussian-based segmentation method for automatic segmentation of speech at syllable-level, the Percentage of Correctness (PoC) prior to manual adjustments is better than the minimum phase GD-based segmentation method [**145**].

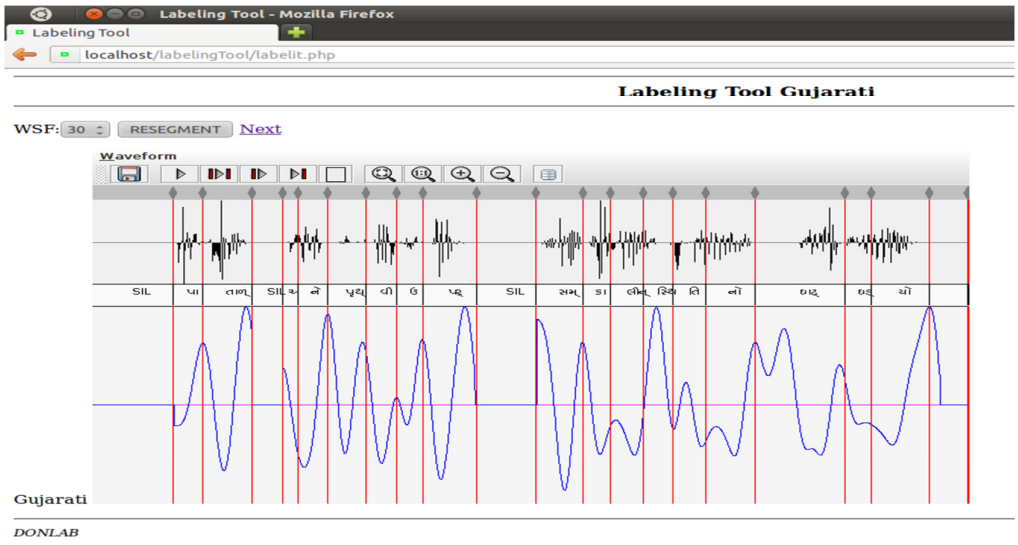


Figure 3.8: The DONLabel labeling tool for Gujarati after manually correcting the labels. After [143**].**

In the USS approach, only the speech sound units that are present in the speech corpus are available for synthesis. To synthesize all possible input text, at least one instance of each syllable must be present in the speech corpus. However, it is practically very difficult to cover all the syllables of a language. As phones are fixed and small in number, if a syllable is not present during synthesis, then syllables are split into phones and phones are concatenated to synthesized speech. In short, a *fallback* mechanism is implemented to build a system that can synthesize every incoming text. The detailed description of the USS system building in other Indian languages using Festival framework [**111**] is presented in [**146**].

3.4.1.3.4 Hidden Markov Model (HMM)-based Speech Synthesis in Gujarati

The HMM-based synthesis framework gives a general setup for context modeling and is easily adapted to other languages [**147**], [**148**]. To build an HTS system, we require a mapping from Gujarati UTF-8 character to *common* roman scripts. To that effect, the mapping was modified for the Gujarati language from [**149**]. For HTS voice building, the text needs to be converted into unicode characters, the sentence is converted into an array of words and words are converted into the syllables. From each syllable pairs, three characters are taken as unicode characters for the Gujarati language having a length of *3* characters. In addition, schwa deletion and anuswara rules are also included.

If an HTS system has to be developed at syllable-level or word-level, then a large number of models are required to model the system. Therefore, for HTS, it is appropriate to build the system at phoneme-level. There are *49* phonemes in the Gujarati language. Hence, we can build system at phoneme-level with very less amount of speech data as compared to syllable-level. We use phoneme as the basic speech sound units for building HTS in the Gujarati language. Phoneme-based labeling was done using forced Viterbi alignment as well as Spectral Transition Measure (STM)-based method [**150**]. Except context-dependent modeling, every block diagram of HTS is language-independent. However, the contextual information is language-dependent [**151**]. The HMM framework provides a general framework for sufficient context modeling that can easily be adapted to other languages. For the phonemic representation of Gujarati language, a set of *49* phonemes were taken that are broadly classified into silence (i.e., SIL), *36* consonants and *12* vowels. In order to build context-dependent HMMs, we require different groups of phonemes. For the Gujarati language, to do classification, we use International Phonetic Alphabet (IPA) chart of Gujarati language for consonants and vowels as in [**152**]. Some examples of classification of phonemes used for question set preparation are:

- Front Vowel: {*-i+*,*-ii+*,*-ee+*,*-ae+*}
- Affricates Consonants: {*-c+*,*-ch+*,*-j+*,*-jh+*}
- Fricatives: {*-ph+*,*-sx+*,*-sh+*,*-s+*,*-h+*}

In this way, several classifications of Gujarati phonemes has been done. In order to build HTS in Gujarati language, *105*-dimensional MFCCs per frame, *3*-dimensional log(F_0) and the *penta-phone* contextual factor is considered [**153**]. Once the HTS is

developed, both subjective and objective evaluations can be considered to test the *naturalness* and speech *intelligibility*.

3.4.1.3.5 Blizzard Challenge 2014 Dataset

The Blizzard Challenge 2014 is entirely dedicated to building TTS voices in Indian languages [**25**]. The six Indian languages included in the challenge were Hindi, Tamil, Telugu, Gujarati, Rajasthani and Assamese. For each language, *2* hours of data sampled at *16* kHz was provided for the challenge. The data was recorded by professional speakers in a high-quality studio environment. Only the speech data and the text were provided in *UTF-8* format. The participants could use any information such as phonesets or labels from other resources. There were *9* participants in the challenge and each was supposed to build one system in each language. Therefore, we have *9* systems for each language under consideration, i.e., from *A* to *I* and *K*. In particular, system *A* consists of natural speech signals, system *C, D, E* and *K* uses HMM-based synthesis technique, *D* used a hybrid approach for Hindi language, systems *G* and *J* use USS-approach and system *F* uses an HMM-DNN approach. The *H* and *I* systems use alternate USS and HTS for *3* languages. The baseline system *B* for each language was build using the speaker-independent HTS-2.2 + STRAIGHT scripts2. The data was labeled at the phone-level using the HMM labeling script (EHMM) in FestVox3 [**25**]. For LTS, a set of simple naive first-order approximations were used for each language. From the Blizzard Challenge 2014 database, we consider two languages, Gujarati and its lexically similar counterpart, i.e., Hindi. The MOS scores for Gujarati and Hindi for the Blizzard challenge submission are given in Figure 3.9. Out of the submissions *B* to *J*, the systems in Gujarati are *C* to *H* and submissions in Hindi are *B* to *H* and *K*. The wavefiles are not available for the remaining systems in the Blizzard database. It is observed that from Figure 3.9 that the almost all systems in Gujarati had a MOS score > *2* (except I), while for Hindi, systems *B*, *H* and *I* had a MOS score from $1 \leq 2$.

Once the database is fixed and features are extracted, it is required to train a classifier and test the accuracy with a standard measure. With respect to this, we discuss the GMM classifier and the performance measures that are used in the task of spoof detection.

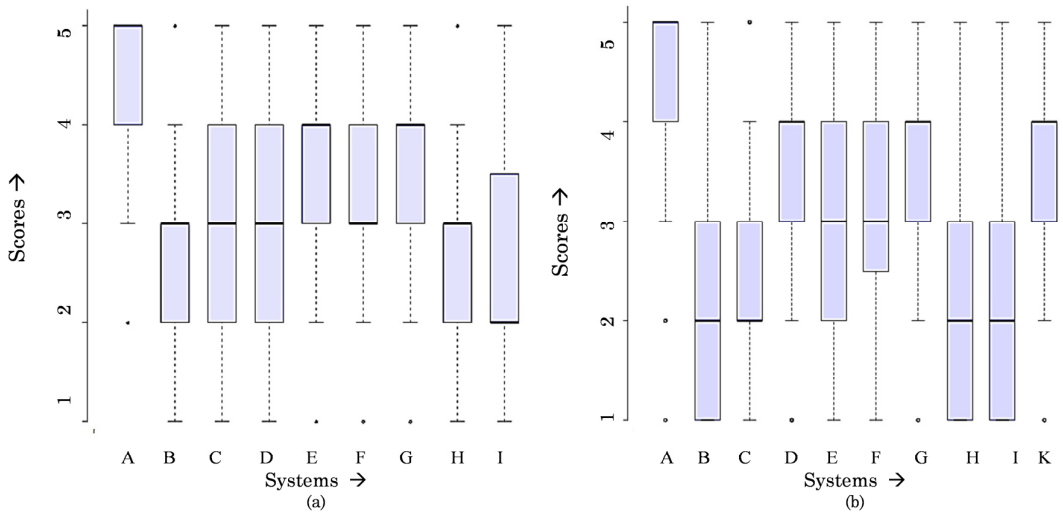


Figure 3.9: The MOS of various systems at the Blizzard Challenge 2014 for (a) Gujarati language and (b) Hindi language. Adapted from [25].

3.4.2 Gaussian Mixture Model-based Classification System

A Gaussian Mixture Model (GMM) is a probability density function represented as a weighted sum of Gaussian component densities. A GMM with M component densities is given by the following equation [**154**],

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \mu_i, \Sigma_i), \tag{3.3}$$

where x is a D-dimensional feature vector, w_i are the mixture weights with $\Sigma w_i=1$ and $g(x \mid \mu_i,\Sigma_i)$, $i=1 ,\ldots, M$, are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form [**154**],

$$g(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right\}, \tag{3.4}$$

where μ_i is the mean vector and Σ_i is the covariance matrix for the i^{th} Gaussian. These parameters are represented by the notation, $\lambda=\{w_i,\mu_i,\Sigma_i\}$. The GMM considers the data as the results of the linear combination of several generative Gaussian models. There are several techniques available for estimating the parameters of a GMM. The most popular and well-established method is an ML estimation. One of the most important ML estimation approach is the Expectation Maximization (EM) algorithm. The EM algorithm uses an E-step (to estimate the distribution of the hidden variable given the data and the current value of parameters) and M-step (to maximize the joint distribution of the data and hidden variable) [**155**]. In this thesis,

we use the GMM to model classes corresponding to natural and spoofed speech (using speech from training dataset). It was observed in Chapter 2 that most of the studies use GMM as a simple two-class classifier for the SSD task. Few approaches have also used SVM [**83**] and DNN-based [**85**] classifiers for classification. It was observed in these studies that using the same feature sets, the GMM-based classification system performed well than other classifiers especially on the evaluation set consisting of unknown attacks.

In this study, the GMM is built on the training set of the ASV spoof challenge dataset. The GMM for natural speech (λ_{nat}) is built using entire training dataset of *3750* genuine (i.e., natural) utterances. Similarly, GMM for spoofed speech (λ_{spoof}) is built with *12625* spoofed training utterances. Final scores on a test sequence Y are represented in terms of Log-Likelihood Ratio (*LLR*) obtained from the likelihood values of natural and spoofed speech model. The decision of the test speech being human or spoofed is based on the *LLR*, i.e.,

$$LLR = \log\ (p\ (Y \mid \lambda_{nat})) - \log\ (p\ (Y \mid \lambda_{spoof})), \tag{3.5}$$

where $p(Y|\lambda_{\text{nat}})$ and $p(Y|\lambda_{\text{spoof}})$ are the likelihood scores from the GMM for the natural speech and spoofed speech, respectively. To utilize possible complementary information in the various proposed features, we use the score-level fusion of various features. Considering two features, the combination at score-level is done as follows:

$$LLk_{\ combine} = (1-\alpha_f)LLk_{feature\ set1} + \alpha_f LLk_{feature\ set2}, \tag{3.6}$$

where $LLk_{feature\ set1}$ and $LLk_{feature\ set2}$ are log-likelihood scores of the feature set 1 and feature set 2, respectively. Parameter α_f decides the weights for score-level fusion. Considering the fusion of three features, the combined likelihood is given as,

$$LLk_{\ combine} = \alpha_1 LLk_{feature\ set1} + \alpha_2 LLk_{feature\ set2} + \alpha_3 LLk_{feature\ set3}. \tag{3.7}$$

The fusion factors in eq. (3.6) and eq. (3.7) are selected such that the sum of all the α*'s* are one, i.e., $\sum_i^3 a_i = 1.$. This will assist to know the relative contribution of the features during the score-level fusion. It is to be noted that in this work, for the training part, the GMM models, i.e., the λ_{nat} and λ_{spoof} are built only once using the training set of the ASV spoof database. However, the testing is done on three sets,

i.e., evaluation set (*S1-S10* spoofs) of the ASV spoof challenge, the Blizzard Challenge 2012 dataset and Blizzard Challenge 2014 dataset.

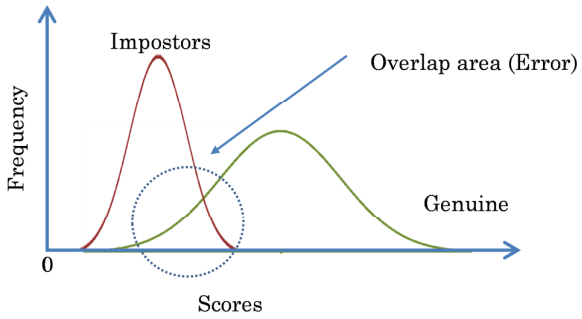


Figure 3.10: An example of likelihood scores for natural and impostor speech.

Figure 3.10 shows an example of possible likelihood score distribution for natural speech and impostor speech. The frequency of the impostor likelihood is more because the number of impostors is generally more as in the case of ASV spoof challenge dataset or can be similar as in Blizzard data. It is observed that impostor scores and genuine scores overlap with each other resulting in error in classification, equivalent to the area under the overlap. The point where the two scores overlap is the threshold score value at which minimum classification error occurs.

Motivation of Score-Level Fusion

In this thesis, we consider score-level fusion instead of feature-level fusion. This is because with the feature-level fusion, the dimension of the resultant feature vector will increase and it was observed to be a strain on the GMM modeling in terms of computation. In addition, it is also difficult to identify the contribution of the individual feature set out of the fused features sets in improving or degrading the performance. However, in the case of score-level fusion, the fusion weight parameter α_f gives the contribution of the feature sets towards the performance of the system. In addition, the feature-level fusion is possible only if the features are extracted from the same duration of the analysis window, which is not the case always. For example, the spectral features are generally extracted from *20-30 ms* of the window while prosodic features are extracted from the larger window of *100-200 ms* to capture the suprasegmental information. Hence, in these cases, the feature-level fusion is not performed and hence, we opted for score-level fusion.

3.4.3 Performance Measures

In evaluating a binary classifier, two types of errors exist, namely, False Acceptance (FA) and False Rejection (FR). A standalone detector system could falsely reject a genuine trial (a false rejection) to the ASV system or falsely accept a spoof or impostor trial (a false acceptance) and allow it to pass through an ASV system. The possible outcomes of an SSD are shown in Table 3.4.

Table 3.4: The confusion matrix of decision trials in an SSD task

Actual Observations (S_n)	Predicted Observations (P_n)		
	Natural	Spoofed	Total
Natural	True Positive (TP)	False Reject (FR)	AP
Spoofed	False Accept (FA)	True Negative (TN)	AN

AP=Actual Positive, AN= Actual Negative

The error rates are expressed as False Acceptance Rate (FAR), i.e., ratio of FA to actual number of positives (natural) and False Rejection Rate (FRR), i.e., ratio of FR to actual number of negatives (spoofed), i.e.,

$$FAR = \frac{FA}{TP + FR} \quad \text{and} \quad FRR = \frac{FR}{FA + TN}. \tag{3.8}$$

The plot of FAR and FRR with respect to the scores ordered in ascending order is shown in Figure 3.11 (a). There is a trade-off between FAR and FRR and both are found to be of equal value around a particular score value (threshold value, i.e., th_{EER}). The value of the error at which the FAR and FRR are equal is known as the Equal Error Rate (EER) which is used as a performance measure [**156**].

Detection Error Tradeoff (DET) Curve

The DET curve is a graphical interpretation of the performance of the classification system for various features using the FRR and the FAR [**157**]. It gives uniform treatment to both FRR and FAR for evaluation of system performance. The DET curve is a plot of FRR on the vertical-axis and the FAR on the horizontal-axis at various threshold levels. In the DET curve, the operating point where FAR and FRR becomes equal is referred to as EER. It serves as a boundary between the output of positive and negative classes. Figure 3.11 (b) shows an example of the DET curve with the interpretation of various regions in the DET curve. The deviation of the operating point from the EER emphasizes any one of the two errors (i.e., FAR or

FRR). In the case of SSD task, the deviation to the vertical-axis provides better security as the FAR is low. However, the FRR will be high which is not intended for ASV systems (will cause user inconvenience as genuine speakers will be rejected). Figure 3.11 are hypothetical plots to provide a pictorial representation of the ideal scores and DET curve. The plots with the real data under consideration will be observed in the following Chapters.

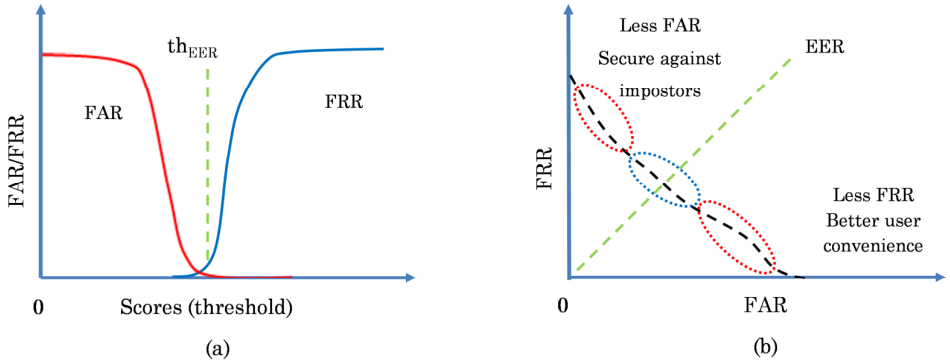


Figure 3.11: (a) The FAR and FRR with respect to the scores ordered in ascending order and (b) the DET plot of FAR *vs.* FRR as varying thresholds [157].

In [**20**], the approach used to estimate the average % EER comprises of estimating the EER individually for each of the spoofing algorithm and then averaging all the EERs. This is shown as Approach 1 in Figure 3.12.

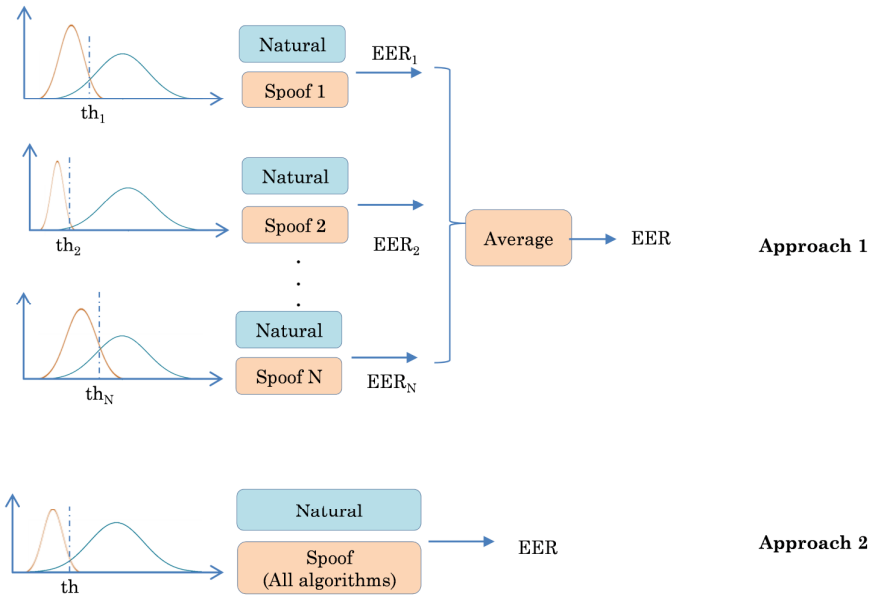


Figure 3.12: Evaluation scheme for computation of the average EER.

However, with such an approach different thresholds are obtained for each of the spoofing algorithms. In the real case for a completely unknown attack, an SSD system should have a fixed threshold below which the speech should be detected as spoofed. Hence, in our study, instead of using EERs estimated from several threshold values, we estimate a universal EER (average EER) assuming only two classes of natural and spoofed speech (as shown by Approach 2 in Figure 3.12). Based on this EER, the individual EERs are obtained by identifying if the likelihood scores for a particular spoof (from a specific spoofing algorithm) was greater than or less the threshold. This can be interpreted as breaking down FAR values for each attack. However, they sum up to the final EER, and hence, the same nomenclature of Individual EER is used in this thesis. The Approach 2 is a much realistic case and it has also been used and accepted as a performance measure in the recent ASV spoof 2017 challenge [**22**]. In fact, the result of pooled EER is also shown in one of the recent works were it has been observed that using one threshold for all detection types, is more realistic for real applications and building a robust spoofing detection system in the real scenario is still a difficult task [**92**].

3.5 Chapter Summary

This Chapter presented the details of the TTS and voice conversion framework. The details of the ASV spoof challenge database and the Blizzard Challenge databases along with the spoofing algorithms are provided. The general architecture of the spoof detection system to be used in this thesis is discussed. In the following Chapters, the various approaches or anti-spoofing techniques based on excitation source, vocal tract as a system (i.e., filter) and source-filter interaction will be presented. Considering the best performing system at the ASV spoof 2015 challenge, in the next Chapter, we first discuss the spectral features or the system-based features proposed in this thesis to identify the cues responsible for unnaturalness in synthetic or converted speech for the SSD task.

Chapter 4.
System-based Features

4.1 Introduction

This Chapter presents the development of system-based features for the Spoofed Speech Detection (SSD) task. It was observed in the previous Chapter that there are several reasons due to which the Synthetic Speech (SS) and Voice Converted Speech (VCS) can sound unnatural. For generating speech an approximate model is used for the spectral representation which cannot capture all the complexities of the vocal tract system. Hence, there will be significant differences between system-based features for natural and machine-generated speech. Moreover, these spectral parameters are extracted and processed framewise while the human speech production is a continuum process. Hence, the variations across frames are also essential. Thus, in this Chapter, we explore the traditional Mel Frequency Cepstral Coefficients (MFCC) and propose novel Cochlear Filter Cepstral Coefficients plus Instantaneous Frequency (CFCCIF) for the SSD task. In addition, Subband Autoencoder (SBAE), i.e., an AE architecture modified to incorporate the human perception mechanism is used for the detecting natural and spoofed speech.

4.2 The Perception Information

For humans, machine-generated speech may sound natural or unnatural based on its quality. Likewise, machines cannot always detect the unnaturalness in the speech signal and hence, may confuse synthetic speech as natural and vice-a-versa. It has been observed that humans outperform detection systems in identifying certain kinds of spoofs [**123**]. Thus, it is essential to exploit and embed the perception mechanism occurring in the ear into the features used for the SSD task. The features (discussed in this Chapter) use some knowledge of the perception mechanism in the ear to generate a feature-level representation. As far as the perception mechanism in the human ear is concerned, the cochlea stands out to be the most vital organ. Brief information of the internal structure of the ear is presented in Figure 4.1 (a).

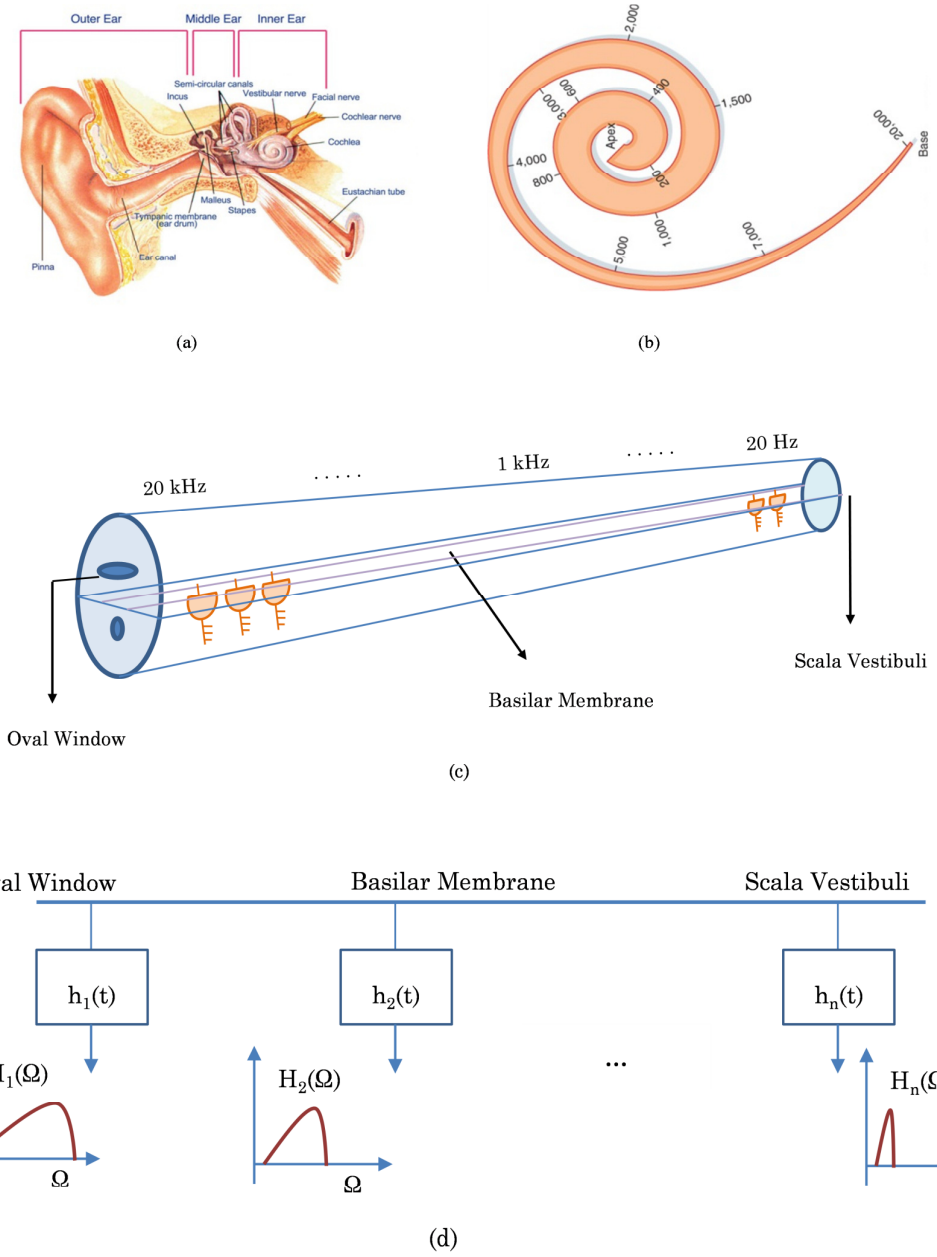


Figure 4.1: (a) Anatomy of the human ear. Adapted from [158], (b) range of frequencies in cochlea (*20* Hz - *20* kHz). Adapted from [159], (c) the uncoiled cochlea. After [21] and (d) the signal processing abstraction of the cochlear filters. After [21].

The ear is divided into the outer, middle and inner ear. The inner ear consists of a coiled snail-like structure called the cochlea. Within the cochlea runs the Basilar Membrane (BM) and along the BM are present several Inner Hair Cells (IHC). Several short hair structures are located on the IHC that deflects when the BM in the human cochlea vibrates at different regions depending on the frequency of the

incoming sounds. Depending on the region of BM and the hair cells that vibrate, different nerves are fired informing the brain about the presence of certain frequencies.

As seen in Figure 4.1 (b), the base region excites to high frequency and gradually decreasing to low frequency towards the apex regions. As the BM vibrates around a particular region for a particular tone of sound, we can interpret human cochlea as bank of subband filters spaced from oval window, i.e., stapes (base) to scala vestibule (apex) whose impulse response are understood to be as that of bandpass frequency response fine tuned to a particular frequency band. A signal processing abstraction to the hair cell excitation to a particular frequency is very well explained in (Chap. 8, pp.402 [**21**]) and demonstrated in Figure 4.1 (d). It can be seen from Figure 4.1 (d) that a particular region on the BM can be represented as a Linear Time-Invariant (LTI) bandpass filter tuned to a particular frequency. This filter has a constant Q, i.e., ratio of the center frequency to the bandwidth of the filter is a constant. Thus, the bandwidth of the filter decreases as the frequency increases from the oval window towards the scala vestibule.

These signal processing abstractions were introduced in MFCC-based features having center frequencies at a Mel scale and symmetrical triangular filterbanks. Ideally, these filters are found to be asymmetric with a steeper response at the right than at the left in (Chap. 8, pp. 398 [**21**]). Similar abstraction has been incorporated in CFCC-based feature representation. In addition to the bandpass representation, the output of each cochlear filter is considered as amplitude and frequency modulated wave. It is known that the envelope of each output of the cochlear filter, its Instantaneous Frequency (IF) and analytic phase are important features used by auditory levels for speech perception (Chap. 8, pp. 403 [**21**]). Earlier in [**160**], an auditory-based distortion measure was used to find the perceptual dissimilarity between speech segments and improve the quality of synthesized speech by selecting speech sound units based on the auditory distortion measures. Next, considering few applications of IF, it has been shown that the short-time IF spectrum contributes to the speech intelligibility as much as the short-time magnitude spectrum [**161**]. In [**162**], the subband IF is used with the envelope from subband filter outputs for speech recognition task. Thus, features derived from analytic phase may be complementary to that of the features derived from magnitude spectrum. Therefore,

in the later Sections of this Chapter, we propose using IF information with the envelope at the output of each subband filters to detect human *vs.* spoofed speech. Prior work shows that countermeasures are designed based on the observation different dynamic variation in the speech parameters of SS and natural speech signal [**97**]. In [**97**], use of frame differences as a discriminative feature was used due to the fact that in the HMM-based speech synthesis, the speech parameter sequence is generated to maximize the output probability and hence, the variation in likelihood will be less as compared to natural speech. In [**99**], higher-order MCEP of SS revealed less variance than that of natural speech signal. This is because, the higher-order MCEP are smoothed during HMM model parameter training and synthesis. Next, as the feature extraction process in SS and VC speech generation is framewise, we use derivative operation to capture transient variations across the frames to assist in the SSD task. The use of derivative enhanced the high frequency regions in the representations which is a possible reason for improved performance. In the CFCCIF features, we propose capturing variations across frames to detect natural *vs.* spoofed speech. The key idea here is that the human speech production system does not produce speech in a frame-by-frame pattern while feature extraction in speech synthesis and voice conversion is generally at frame-level. Thus, using a derivative operation to capture variations across features will assist in the SSD task.

In addition, the subband processing has also been incorporated in AE framework to develop SBAE architecture for spoof detection task. It was observed that the SBAE features learned the variations across frames which are generally more in machine-generated speech than in normal speech which resulted in better spoof detection. The human perception mechanism is a highly complex process and incorporating these in the development of features will aid in extracting the features that match more closely with the perception for hearing.

4.3 Mel Frequency Cepstral Coefficients (MFCC)

The MFCC introduced by Davis and Mermelstein in the 1980's [**8**] has been widely used since then as state-of-the-art features in several speech processing applications such as speech recognition, speaker and language recognition, speech synthesis, emotion recognition, etc. Speech signal represents the continuously changing movement in the vocal tract system over a short time and thus, depending on the

type of information needed, the speech signal is framed and processed. The brief representation of the feature extraction process in MFCC is shown in Figure 4.2.

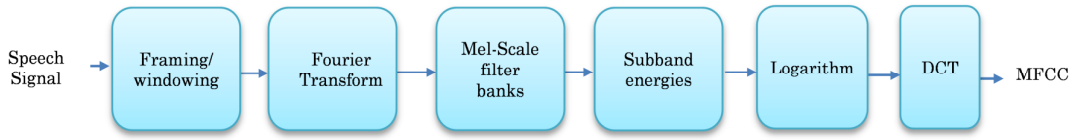


Figure 4.2: Schematic diagram of the MFCC feature extraction process. After [8].

Once the speech is framed, the next step is to calculate the power spectrum of each frame. The resulting power spectrum with varying frequencies is passed through various subband filters as in the human ear. These frequencies are centered according to the Mel scale. The Mel scale relates perceived frequency, or pitch (which is a perceived pitch, i.e., a perception phenomenon) [**21**], [**163**], of a pure tone to its actual measured frequency. It takes a set or range of frequencies (*f*) at linear scale and converts it into a logarithmic scale by the relation [**164**],

$$M(f) = 2595 \log_{10}(1 + f / 700). \tag{4.1}$$

The Mel filterbank is a set of symmetric triangular subband filters. As the frequencies get higher, the subband filters are wider. Thus, the power spectrum is passed through several Mel subbands to get filterbank energies. In MFCC or any other spectral features, the subband energy is computed. Once filterbank energies are obtained, the logarithm is taken. The final step is to compute the DCT of the log filterbank energies. As the filterbanks are all overlapping, they are correlated with each other. The DCT decorrelates the energies and allows choosing few initial values of the DCT known as static MFCC features for speech processing applications.

The MFCC feature vector describes only the power spectral envelope of a single frame. However, speech is known to have information present in the dynamics of the trajectories of the MFCC over time. Using (or appending) the delta (velocity), i.e., the dynamic information of the MFCC feature vector of each frame along the static features, is known to capture better information than static alone. However, this may vary with the application. The delta-delta (acceleration) coefficients are also used as additional dynamic features along with the static and the delta features.

4.4 Cochlear Filter Cepstral Coefficients (CFCC)

The subband in Mel filterbank is a triangularly-shaped symmetric filter. However, it has been shown that the filters in the inner ear called as a cochlear filter are rather asymmetric (Chap. 8, pp. 398 [**21**]). Therefore, the use of auditory-based filters in place of triangular filters will aid in capturing additional perceptual information as compared to MFCC. With respect to this, we discuss the Auditory Transform (AT) developed in [**165**] which is a set of subband filters designed in such a way so as to depict the hearing mechanisms in the ear. The parameter extraction procedure for auditory-based cepstral coefficients, consists of series of cochlear filterbank based on the AT, the hair cell function of the BM, nerve spike density representation, nonlinearity (loudness function) and DCT [**166**] as shown in Figure 4.3. The CFCC features had been used for robust speaker identification [**166**], classification of fricatives sounds [**167**], etc. Following sub-Section describes briefly the AT and the procedure to estimate CFCC features.

Figure 4.3: Schematic diagram of the CFCC feature extraction process. After [166**].**

4.4.1 Auditory Transform (AT)

The AT has well defined wavelet properties with a well-defined inverse transform [**165**]. It converts the time-domain signal into a set of filterbank outputs with frequency responses similar to those in the BM of the cochlea.

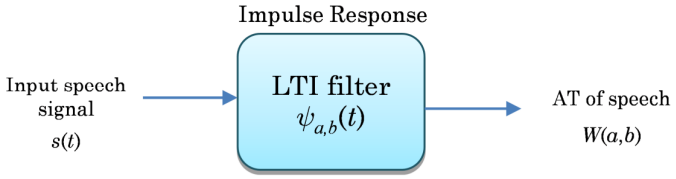


Figure 4.4: Auditory transform as LTI filtering of speech. After [168**].**

The AT process can be considered as a LTI filtering process as in Figure 4.4. Let $s(t)$ be the speech signal and cochlear filter be $\psi(t)$ then the AT of $s(t)$ (i.e., $W(a,b)$), w.r.t. $\psi(t)$ as the impulse response of BM in the cochlea is defined as [**165**]- [**166**].

$$W(a,b) = s(t) * \psi_{a,b}(t), \tag{4.2}$$

$$W(a,b) = \int_{-\infty}^{+\infty} s(\tau)\psi_{a,b}^{*}(t-\tau)d\tau, \tag{4.3}$$

where $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$. In eq. (4.2), * indicates convolution operation, $a \in \mathrm{R}^{+}$ and $b \in \mathrm{R}$, $s(t)$ and $\psi(t)$ belongs to Hilbert space $L^2(\mathrm{R})$ (i.e., space of finite energy signal) and $W(a,b)$ represents traveling waves in the BM. The factor 'a' is the scale or dilation parameter, which allows changing the center frequency (f_c) while factor 'b' is the time shift or translation parameter. It should be known that the f_c of the cochlear filter is also called as the Characteristic Frequency (CF) [**163**]. The energy remains equal for all a and b. Therefore,

$$\int_{-\infty}^{+\infty} |\psi_{a,b}(t)|^2 = \int_{-\infty}^{+\infty} |\psi(t)|^2 dt. \tag{4.4}$$

The cochlear filter is defined as [**166**],

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\left(\frac{t-b}{a}\right)^{\alpha} \exp\left[-2\pi f_L \beta\left(\frac{t-b}{a}\right)\right] \times \cos\left[2\pi f_L\left(\frac{t-b}{a}\right) + \theta\right] u(t-b). \tag{4.5}$$

Parameters α and β determine the *shape* and *width* of cochlear filter, respectively and θ is selected such that the following *admissibility* condition for mother wavelet (i.e., $\psi(t)$) is satisfied [**168**]:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \Rightarrow \psi(\omega)\big|_{\omega=0} = 0. \tag{4.6}$$

Thus, $\exists$ a number C_ψ such that, $C_\psi = \int_0^{+\infty} \frac{|\psi(\omega)|^2}{\omega} d\omega < \infty$ (Theorem 4.3, pp. 81, [**168**]).

This means that the mother wavelet $\psi(t)$ is a *bandpass* filter. That is, wavelet $\psi(t)$ has a zero dc value and hence, it is a *bandpass* filter. The value of a can be derived from the center frequency (f_c) and lowest frequency (f_L) of the cochlear filterbank, i.e.,

$$a = \frac{f_L}{f_c}. \tag{4.7}$$

For the i^{th} subband filter, its corresponding value of a, i.e., $\{a_i\}$ is pre-calculated for the required f_c of the cochlear subband filters at band number $i \in [1, N_F]$, where N_F is the total number of subband filters. An example of *14* filters with α=3 and for β=*0.035* and β=*0.35* for different value f_c equally space on a linear scale till *8* kHz is

shown in Figure 4.5. It is observed from the linear scale that the bandwidth of the filter is high at the higher frequencies than at the lower frequencies. The asymmetricity of the cochlear filter can be viewed much better in the log-scale than in the linear scale.

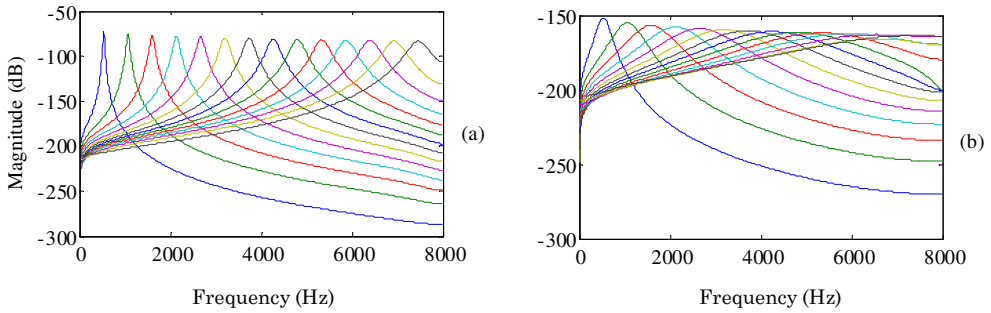


Figure 4.5: Responses of *14* cochlear subband filters on a linear scale with (a) α=3 and β=0.035 and (b) α=3 and β=0.35. After [166].

4.4.2 Other Operations

Once filtering process is done by the cochlea in the ear, the IHC acts as a transducer for the movements of BM. As motion of the hair cell is only in the *positive* direction, the following nonlinear function of hair cell describes this motion [**166**], i.e.,

$$h(a,b) = (W(a,b))^2; \qquad \forall W(a,b), \tag{4.8}$$

where $W(a,b)$ is the filterbank output or a subband signal. The hair cell output of each filterbank is converted into a representation of the nerve spike density, i.e.,

$$S(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} h(i,b), \qquad l = 1, L, 2L, ...; \ \forall i, j\ , \tag{4.9}$$

where d is the window length, i is the i^{th} subband, j is the frame count and L is the window shift duration. The output of the above is further applied for scales of loudness functions such as logarithm or *cubic root* nonlinearity. Finally, DCT is applied to decorrelate the features.

4.5 Cochlear Filter Cepstral Coefficients plus Instantaneous Frequency (CFCCIF)

4.5.1 Procedure of Extraction of CFCCIF

This Section presents the proposed variant of CFCC features that uses both the

envelope and average IF from the subbands to give the Cochlear Filter Cepstral Coefficients plus Instantaneous Frequency (CFCCIF) features.

4.5.1.1 Average Instantaneous Frequency (AIF) Estimation

In CFCC, the nerve spike density performs averaging operation on each subband signal which in turn removes the Temporal Fine Structure (TFS) or fast temporal modulations as shown in Figure 4.6 (c) [**169**]. Hence, CFCC does not incorporate TFS information. Furthermore, at every CF of the cochlear filter (i.e., the center frequency of the cochlear filter, f_c as in eq. (4.7)), rapid phase shift of the travelling wave occurs at every f_c from base to apex [**163**]. We believe that this rapid change is being captured by the derivative of instantaneous (analytic) phase (which is referred to as IF) of corresponding subband signal. For the SSD task, in vocoder-dependent spoofs, the phase information is generally lost. On the other hand, for vocoder-independent speech, the phase mismatch and temporal discontinuity occur due to joining or concatenation of the speech sound units. Hence, we propose to use IF for every subband along with the envelope representation obtained in the CFCC framework for the SSD task.

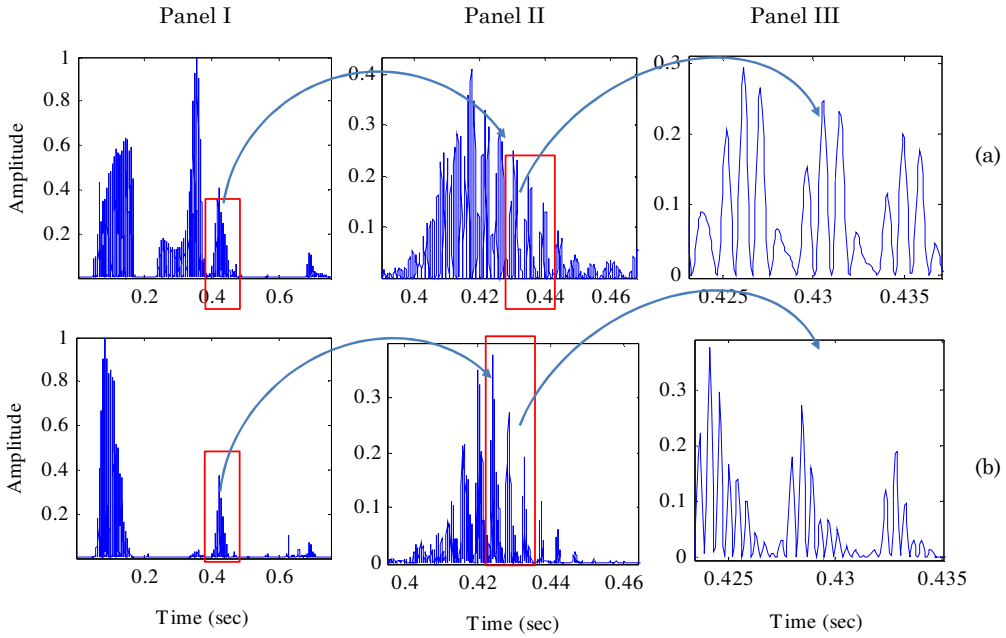


Figure 4.6: For a subband around (a) f_c=550 Hz and (b) f_c=1100 Hz, Panel I: the slow modulations that roughly correlate with the different syllable length segments of the utterance, Panel II: modulations due to interharmonic interactions occur at a rate that reflects the fundamental frequency (F_0) of the signal and Panel III: fast temporal modulations due to the frequency component driving this subband around f_c. After [169].

Hence, for the IF estimation, let $x_i(t)$ be the signal for the i^{th} subband. For the real signal $x_i(t)$, its complex *analytic* representation is given by,

$$x_{a_i}(t) = x_i(t) + jx_{h_i}(t), \tag{4.10}$$

where $x_{h_i}(t)$ is the Hilbert transform of the signal $x_i(t)$, given by the inverse Fourier transform of $X_{h_i}(\omega)$, where,

$$X_{h_i}(\omega) = \begin{cases} +jX_i(\omega) & \omega < 0 \\ -jX_i(\omega) & \omega > 0 \end{cases}. \tag{4.11}$$

Thus, the amplitude (Hilbert) envelope of $x_i(t)$ and the instantaneous phase for the i^{th} subband is given as,

$$|x_{a_i}(t)| = \sqrt{x_i^2(t) + x_{h_i}^2(t)} \text{ and } \phi_i(t) = \tan^{-1}\left(x_{h_i}(t) \Big/ x_i(t)\right). \tag{4.12}$$

Therefore, for the i^{th} subband, the IF derived from derivative of *unwrapped* instantaneous phase $\phi_i(t)$, is given as:

$$IF_i = \frac{d}{dt}\left(\phi_i(t)\right). \tag{4.13}$$

Next, similar to nerve spike density estimation, the framewise average IF for each i^{th} subband is obtained as,

$$AIF(i,j) = \frac{1}{d}\sum_{b=l}^{l+d-1} IF_i(b), \quad l = 1, L, 2L, ...; \ \forall i, j, \tag{4.14}$$

where d is the window length, j is the frame count and L is the window shift duration.

4.5.1.2 The CFCCIF and CFCCIFS Representation

For each subband, the envelope (estimated in eq. (4.9)) needs to be combined with the average IF (estimate in eq. (4.14)). In particular, for each of the i^{th} subband, using eq. (4.9) and eq. (4.14), we have,

$$z(i,t) = S(i,t) \cdot AIF(i,t), \tag{4.15}$$

where $z(i,t)$ is the representation obtained after multiplying subband envelope and average IF features for the i^{th} subband. In [**162**], the subband IF was used explicitly by concatenating it with the envelope from subband filter outputs for the task of speech recognition. However, with this, the feature dimension increases to twice. In [**170**], multiplication of envelope and fine structures estimated from the Hilbert

transform of bandpass filtered signal was carried out. This was done to investigate the relative perceptual importance by Chimaera synthesis [**170**]. Therefore, we use multiplication of both envelope and average IF features to preserve the relevant information at the same feature dimension. Further, the multiplication operation will suppress the random IF estimated in silence regions by the low amplitude values of the envelope structure.

4.5.1.3 The Difference Operation

As in traditional MFCC and CFCC operation, the output of the filterbank is applied for nonlinearity (logarithmic) operation followed by DCT to obtain feature representation. Instead of directly using the representation obtained by multiplying the $S(i,j)$ and $AIF(i,j)$, the change across frames is computed via a derivative operation. Thus, differentiating eq. (4.15) partially w.r.t t on both the sides, we get,

$$\therefore \frac{\partial(z(i,t))}{\partial t} = AIF(i,t)\frac{\partial S(i,t)}{\partial t} + S(i,t)\frac{\partial AIF(i,t)}{\partial t}, \tag{4.16}$$

Thus, the derivative of $z(i,t)$ representation is the sum of changes in nerve spike density weighted by average IF and the changes in average IF weighted by the nerve spike density. The efficiency of the proposed features lies in exploiting the dynamic information via derivative operation for the present problem of SSD task. The basic idea is that the human speech production is a *continuum* process, with relatively fewer changes in the amplitude and frequency across any speech sound unit (unless there is an abrupt transition from one speech sound unit to another). However, in generating either the SS or the VCS, the features are extracted and processed framewise. Thus, to capture the transient information, the *change* in the envelope (by CFCC) and average IF between consecutive frames is estimated through the derivative operation. Therefore, we explore the transient information by simple one-point derivative (backward and forward difference) and by the symmetric difference. It was observed that both backward and forward derivative gave similar information and hence, only the backward difference is considered hereafter. In the present work, to capture both past and future context information at a particular time instant, we use the *symmetric* difference to estimate the change in CFCCIF features across the frames. The symmetric difference is used as follows,

$$\therefore \frac{\partial(z(i,t))}{\partial t} \approx \frac{z[i,n+1]-z[i,n-1]}{2}, \tag{4.17}$$

where $z(i,t)$ is given as per eq. (4.15). Using symmetric difference which uses both past and present sample amplitude and frequency information smoothens out abruptness due to one sample differentiation and also intuitively represents the fact that both past and future context information are essential for perceiving the transient information at a particular instant. The significance of such context was also observed for syllable boundary detection using STM and the listening test [**171**].

4.5.1.4 Other Operations

Next, the derivative operation is followed by considering the absolute value followed by the logarithm nonlinearity (this is repeated for all the subbands, i.e., $i \in [1,28]$). Finally, DCT is applied framewise. These features obtained using backward difference are known as CFCCIF features and those obtained with a symmetric difference are known as CFCCIFS features. The complete architecture of the CFCC and CFCCIFS features is shown in Figure 4.7.

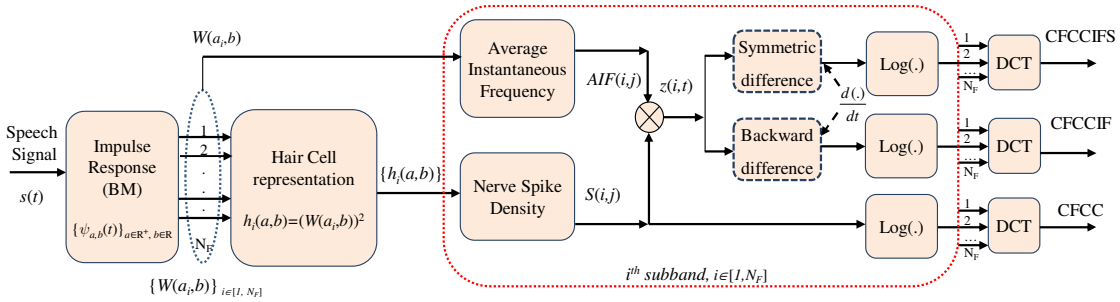


Figure 4.7: Schematic diagram of the CFCC, CFCCIF and proposed CFCCIFS feature extraction process. (Adapted from [91]).

4.5.2 Effectiveness of Derivative Operation

The efficiency of the proposed features lies in exploiting the dynamic information via derivative operation for the present problem of natural *vs.* spoof speech detection. The type of dynamic information that is captured depends on the shape of the cochlear filter shape. Therefore, we study the effect of the envelope and the phase features using a case of a wide bandwidth with $\alpha=3$ and $\beta=0.35$ and a narrow bandwidth cochlear filter with $\alpha=3$ and $\beta=0.035$. Figure 4.8 shows the case for a wide bandwidth cochlear filter. In particular, Panel I, Panel II, Panel III and Panel

IV considers the analysis of natural speech, vocoder-based SS and vocoder-based VCS and unit-selection-based speech (MARY TTS), respectively, with the same text material. The vocoder-based SS and VCS correspond to the utterances from the *S3* and *S7* spoofs of the SAS database [**19**], respectively. Figure 4.8 (a) shows the speech waveform, Figure 4.8 (b) and Figure 4.8 (c) shows the subband energy representation of MFCC and CFCC, respectively. Figure 4.8 (d) shows subband energy representation obtained on multiplying envelope and average IF (i.e., without the derivative operation as in eq. (4.15)), Figure 4.8 (e) and Figure 4.8 (f) shows the subband energy representation of CFCCIF and CFCCIFS using backward and symmetric difference operation, respectively.

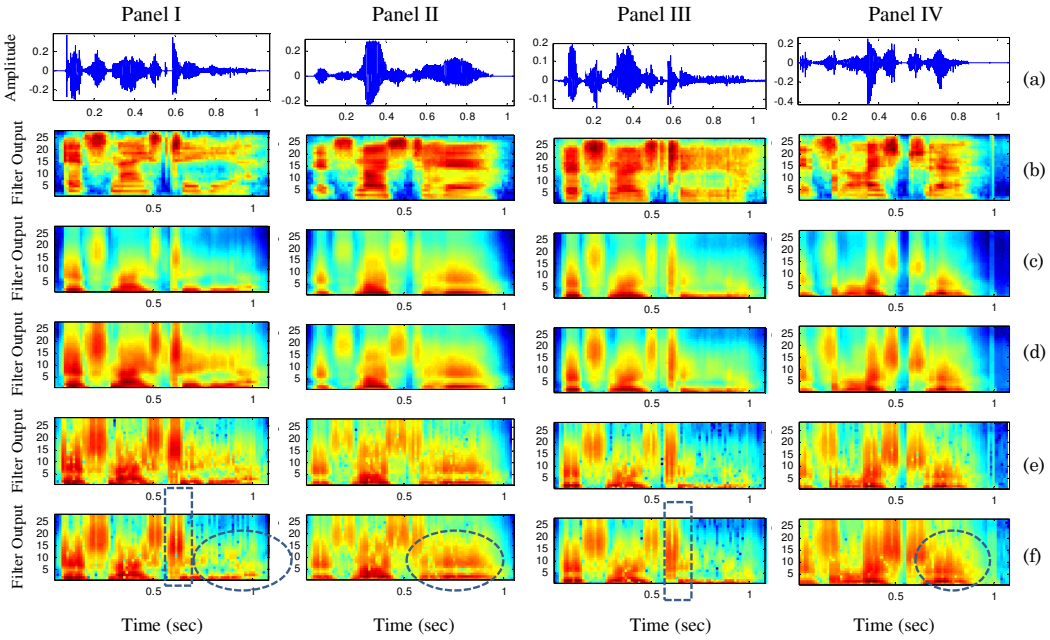


Figure 4.8: Panel I: Natural speech, Panel II: vocoder-based SS, Panel III: vocoder-based VCS and Panel IV: USS-based MARY TTS: (a) speech signal waveform of the utterance /It's nice to hear/ from the SAS database [19], the subband energy representation of (b) MFCC (c) CFCC (d) multiplication of CFCC and IF (without the derivative operation) (e) CFCCIF (using one-point backward derivative) [81] and (f) CFCCIFS (using symmetric difference operation). The cochlear filter parameters are *α=3* and *β=0.35*. Dotted regions show differences in natural and the spoofed speech signal.

As observed in Figure 4.8 (b) and Figure 4.8 (c), there exists fewer pitch (i.e., fundamental frequency F_0) harmonics (due to wider bandwidth) and computational noise (due to filtering of the entire signal rather than estimating FFT frame wise) in the spectra from the auditory transform (CFCC) than MFCC representation [**166**]. The representation in Figure 4.8 (d) obtained by multiplication of envelope and

average IF features (without the derivative operation as in eq. (4.15)) is different from Figure 4.8 (c) is the sense that the high-frequency regions are enhanced by CFCCIF features. It is observed in Panel I and Panel III that VCS is observed to match the characteristics of the natural *speaker* more than SS (the speaker-specific information is retained especially in high-frequency regions [**172**]). However, it is necessary to bring out the information about natural and spoofed speech. Next, the proposed CFCCIF features are shown in Figure 4.8 (e) of all the panels [**81**]. After taking the derivative, features corresponding to natural and spoofed speech have been visible. As the natural speech production mechanism is a continuum and almost constant process at least for a particular speech sound unit, taking the derivative will minimize the energy (as shown by oval regions in Panel I). On the other hand, for SS speech in the similar region, the energy is not minimized even after derivative operation (as shown by oval regions in Panel II), i.e., the speech synthesis generation process was rather discontinuous as compared to the natural speech signal. For VCS, such direct differences were not observed. However, for VCS, the energy intensity is less (as shown by dotted squares in Panel III) than that of the natural speech throughout the speech utterance (as shown by dotted squares in Panel I). Similar inferences were observed for few other utterances of the SAS database [**19**]. Thus, by using symmetric difference, the subband energy representation becomes smoother and in fact, the difference between natural and spoofed speech (i.e., SS and VCS) is much more prominent.

Next, we study the effect of the envelope and the average IF features using a narrow *-3* dB bandwidth (i.e., high quality (Q) factor) cochlear filter with $\alpha=3$ and $\beta=0.035$. Narrower filters are needed for efficient IF estimation (in order to avoid ambiguity in IF estimation). Moreover, in the early auditory processing model of Shamma [**163**], [**169**], high quality cochlear subband filter responds only to frequencies near the center frequencies, and hence, are found to produce more regular (i.e., periodic) synchronized responses even independent of input stimuli (such as noise, harmonic sequence or impulse) [**163**]. The spectrum of the AT is known to preserve the formant information with fewer pitch harmonics (i.e., F_0) and computational noise [**166**]. As observed in Figure 4.9 (c), the formant characteristics are enhanced more in natural speech than the vocoder-based SS and VCS.

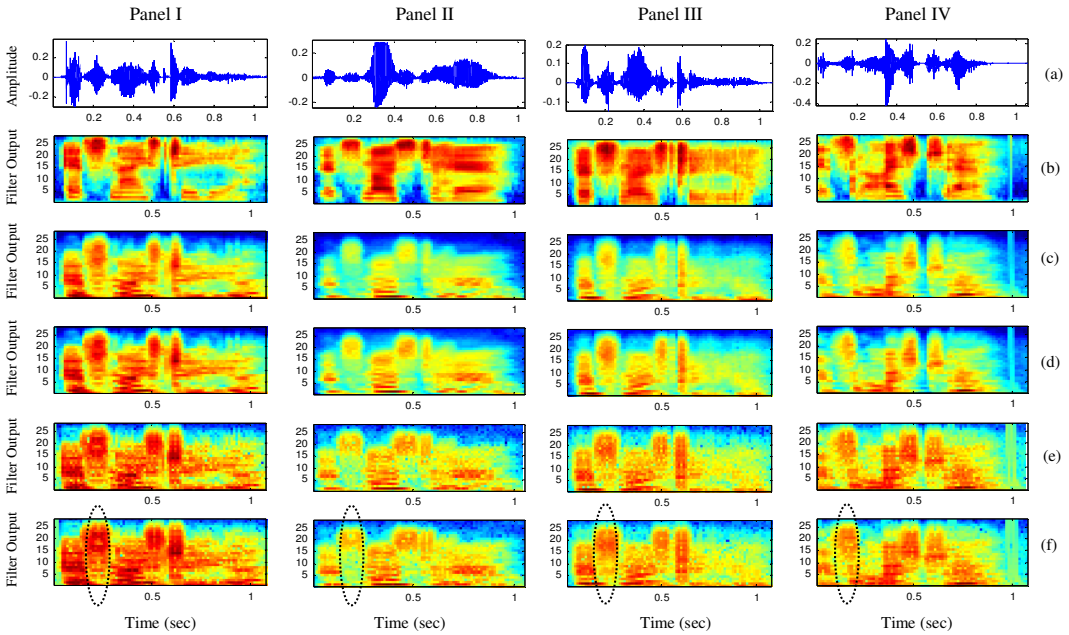


Figure 4.9: Panel I: Natural speech, Panel II: vocoder-based SS, Panel III: vocoder-based VCS and Panel IV: USS-based MARY TTS: (a) speech signal waveform of the utterance /It's nice to hear/ from the SAS database [19], the subband energy representation of (b) MFCC (c) CFCC (d) multiplication of CFCC and IF (without the derivative operation) (e) CFCCIF (using one-point backward derivative) [81] and (f) CFCCIFS (using symmetric difference operation). The cochlear filter parameters are *α=3* and *β=0.035*. Dotted regions show differences in natural and the spoofed speech signal. (Adapted from [91]).

The higher formants are attributes of the natural speech and it is difficult to incorporate it in the machine-generated speech. In the case of USS-based speech, the formant information is intact due to the concatenation of natural speech sound units. However, this depends on the sound units that are picked for concatenation. In Figure 4.9 (d), the representation obtained by multiplication of envelope and average IF features (without derivative) is similar to Figure 4.9 (c) except that the frequency regions are enhanced due to the embedded IF information. For spoofed speech in Panel II and Panel III, the high frequency regions are enhanced after multiplication with average IF. Next, the CFCCIF filterbank representation as in [**81**] is shown in Figure 4.9 (e) for all the panels. After taking derivative, the features corresponding to natural and spoofed speech have been more discriminative. For natural speech, the dynamic variations of the envelope and average IF across the frames were more visible along all the filterbanks as compared to that of the SS and VCS. Furthermore, by using symmetric difference, the subband energy representation is smoother and in fact, the difference between natural and spoofed speech (i.e., SS and

VC) is much more prominent. Therefore, from embedding the average IF information to taking the derivative, the high frequency regions have significantly enhanced as shown by dotted regions in Figure 4.9. In a very recent study, it has been observed that high-frequency regions indeed are essential for spoof detection [**83**], [**84**]. In the case of natural speech, the energy variations are more and along the entire utterance. This is not the case for spoofed speech, especially for vocoder-based speech. The dotted squares show a region of sound \s\ for all the utterances. The envelope and IF representation are significant in this area as compared to the CFCCIF representation in Figure 4.9 (d) without the derivative operation. The use of symmetric difference has considered both the past and the future samples for the derivative operation. This not only makes the representation smoother, but also intuitively represents the fact that both past and future context information are essential for perceiving the transient information at a particular instant.

4.5.3 Experimental Results

4.5.3.1 Parameterization

The MFCC, CFCC, CFCCIF and CFCCIFS features are extracted from *25 ms* of the frame with a shift of *50* % between frames. Both static (s) (without 0^{th} energy coefficient) and dynamic features, i.e., delta (Δ) and delta-delta (ΔΔ) for all the feature sets are extracted. Thus, three different dimensions (*D*) of the feature vector, i.e., *D1*: *12-D* static features, *D2*: *24-D* (*12* static + *12*-Δ), *D3*: *36-D* (*12* static + *12*-Δ + *12*-ΔΔ) are considered. In addition, to estimate the dynamic features, various analysis window intervals are considered for the derivative operation to know the best possible window size for SSD task. In addition, the individual contribution of the dynamic features is also analyzed. The parameters of CFCC, CFCCIF and CFCCIFS are fixed to $\alpha=3$ and $\beta=0.035$ through experiments. These values of α and β give a narrow shape to the auditory filters which help to capture better spoof-specific features. In the next sub-Section, the experimental results on the ASVspoof 2015 challenge database and Blizzard 2012 and 2014 databases are demonstrated. The results of the ASV spoof 2015 challenge database are presented in [**91**]. Details of the GMM-based classification system and the performance measures to be used were described in Chapter 3.

4.5.3.2 Results on the Development Set of ASVspoof challenge Database

Choice of number of subband filters: The experiments related to the choice of a number of subbands for Mel filterbank and cochlear filterbank are presented here. As observed in Figure 4.10 that overall the % EER decreases from *D1* to *D3* feature vector. The MFCC feature set gave high % EER for less number of subband filters as compared to the CFCC, CFCCIF and CFCCIFS. The % EER of the features do not vary much after *25* subband filters especially for *D3* feature vector. Thus, instead of using a large number of subband filters, we use slightly greater than *25,* i.e., *28* subband filters for all the four feature sets.

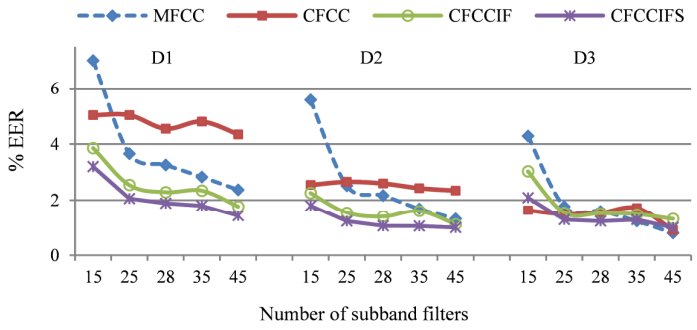
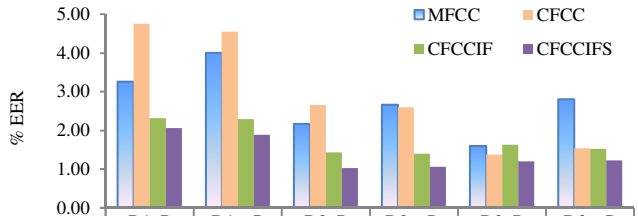


Figure 4.10: Effect of a various number of subband filters on the % EER for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using the *D1*, *D2* and *D3* feature vectors.

The effectiveness of pre-emphasis on speech signal: To study the dependence of the features on pre-emphasizing the speech signal, the % EER was obtained with pre-emphasis (*P*) and using *no* pre-emphasis (*nP*) for MFCC, CFCC, CFCCIF and CFCCIFS features sets as shown in Figure 4.11. The MFCC features have a sensitive dependence to pre-emphasis, i.e., for *nP*, its % EER increases significantly for all sets of feature vectors. On the other hand, the % EER of CFCC-based features (with *P* or *nP*) are almost constant for all feature sets. In fact, on an average, CFCC, CFCCIF and CFCCIFS feature sets perform better without pre-emphasis. Thus, the performance of the cochlear filter-based features is not significantly dependent on pre-emphasis. This is due to embedded bandpass filtering (i.e., due to *admissibility* condition of the cochlear filter, i.e., mother wavelet function $\psi(t)$ as in eq. (4.6) [**81**]).

Thus, for all the experiments, MFCC is used with pre-emphasis filter (i.e., $1\text{-}0.97z^{-1}$) and cochlear filter-based features are used without explicit use of pre-emphasis filter.



	D1_P	D1_nP	D2_P	D2_nP	D3_P	D3_nP
MFCC	3.26	4.00	2.17	2.66	1.60	2.80
CFCC	4.75	4.55	2.66	2.60	1.37	1.54
CFCCIF	2.32	2.29	1.43	1.40	1.63	1.52
CFCCIFS	2.06	1.89	1.03	1.06	1.20	1.23

Figure 4.11: Effect of pre-emphasis on % EER for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using the *D1*, *D2* and *D3* feature vectors (*P*=pre-emphasis with a pre-emphasis factor of (a_{pre}=0.97) and *nP*=no pre-emphasis on speech signal).

Effectiveness of derivative operation in CFCCIF: As discussed in Section 4.5.1, the derivative operation (shown by dashed region in Figure 4.7) before taking log and DCT indeed facilitates the SSD task. As observed in Table 4.1, the % EER is less with the derivative operation. In fact, using derivative in time-domain gives less % EER with just *D2* (static+Δ) feature vector. For higher-dimensional feature vector (*D3*), no difference is found in % EER with and without derivative. The difference is not observed in the EER, however, the likelihood scores are different and the changes/improvements are observed during fusion. Also for consistency across all the feature sets, we use *D3* feature vector.

Table 4.1: EER (in %) for CFCCIF feature set with and without derivative

	Feature Vector		
	D1	**D2**	**D3**
Without Derivative	4.318	2.4878	1.5156
With Derivative	2.287	1.4012	1.5156

Effect of window size to estimate dynamic features: A recent work in [**83**], has shown that the dynamic features alone can contribute effectively to the detection process and achieve almost similar or less % EER than the static features. Along similar lines, Table 4.2 shows the individual contribution of *12-D* Δ and *12-D* ΔΔ features and combined effect of the *24-D* $\Delta+\Delta^2$ features for all the four feature sets considered in the present work. The dynamic features are estimated for four analysis frames ($2n_0+1$), i.e., with $n_0=1$ ($w1$), with $n_0=2$ ($w2$), with $n_0=3$ ($w3$) and with $n_0=4$ ($w4$) corresponding to *18.75 ms*, *31.25 ms*, *43. 75 ms* and *68.25 ms*, respectively, for a F_s of *16 kHz*. It is observed that the dynamic features alone are effective for MFCC

and CFCC only when a *w1* frame window is used. Likewise, CFCCIF and CFCCIFS perform well for the *w2* frame window. For larger frame window (i.e., *w3* and *w4*), the % EER increases significantly. This change is rather more for MFCC and CFCC features than that of the CFCCIF and CFCCIFS features as evident from the average values. Thus, it is observed that the performance of MFCC and CFCC feature sets are more dependent on the window size as compared to the CFCCIF and CFCCIFS feature sets.

Table 4.2: EER (in %) for *12-D* Δ, *12-D* ΔΔ and *24-D* Δ+Δ² feature vectors for all feature sets

No. of Frames	Feature Sets											
	MFCC			CFCC			CFCCIF			CFCCIFS		
	Δ	Δ²	Δ+Δ²	Δ	Δ²	Δ+Δ²	Δ	Δ²	Δ+Δ²	Δ	Δ²	Δ+Δ²
w1	**3.77**	**4.69**	**3.09**	**1.54**	**0.77**	**1.00**	**2.69**	5.78	**1.77**	2.37	5.49	**1.40**
w2	5.83	6.23	5.72	5.18	3.75	4.55	2.83	**3.17**	2.12	**2.23**	**2.80**	1.54
w3	7.09	7.35	7.15	8.98	7.58	8.15	4.38	4.23	3.63	3.03	2.86	2.40
w4	8.15	8.49	8.84	11.41	10.21	10.69	7.06	5.78	5.63	4.58	4.03	3.66
Average	6.21	6.69	6.20	6.78	5.58	6.10	4.24	4.7	3.29	3.05	3.80	**2.25**

It is observed from Table 4.2 that with the Δ + Δ² features used together, the % EER improves than using dynamic features alone. The complementary information in Δ and Δ² features were added when used jointly. For the cochlear filter-based features, namely, CFCC, CFCCIF and CFCCIFS features, the performance was even better than the static features. From Table 4.2, it was observed that the % EER for only Δ² features for CFCCIF and CFCCIFS features were more than *5* % with a *w1* frame window. However, when combined along with their Δ features, the % EER went down to *1.77* % and *1.40* % for CFCCIF and CFCCIFS, respectively. For CFCC with Δ² features, the EER is as low as *0.77* %, however, this is not consistent across window lengths and hence not considered as the best representation. From Table 4.2, it can be concluded that *w1* is the best window for all the features to obtain relatively least % EER. Using this case, the combined effect of static, Δ and ΔΔ features are studied as shown by shaded cells in Table 4.3. It is observed that with the *36-D* feature vector (i.e., static+Δ+Δ²), the % EER of MFCC reduced to *1.6* % as compared to using only static or only dynamic features alone. The CFCCIF and CFCCIFS gave least % EER with *D2* feature vector due to reasons discussed in Table 4.1.

Results of score-level fusion: The fusion of MFCC with either CFCC or CFCCIF or with CFCCIFS is considered and shown in Table 4.3. It is observed that the best % EER on the development set is obtained with a fusion weight, α_f = *0.4* for CFCC

and $\alpha_f = 0.6$ for CFCCIF and CFCCIFS. Thus, CFCCIF and CFCCIFS feature set on score-level fusion added more *complementary* information in reducing the % EER than MFCC alone. The fusion of proposed CFCCIFS and MFCC using *D3* feature vector gave the least % EER of *0.66* amongst all combinations. In fact, scores obtained from MFCC and CFCCIF with α_f=*0.6*, i.e., with an EER of *0.83* % was submitted at the ASVspoof 2015 challenge which was found to be relatively the best performing system among all the *16* submissions [**20**]. Thus, the MFCC-CFCCIFS system performs much better detection than the MFCC-CFCCIF-based SSD system.

Table 4.3: EER (in %) for score-level fusion of MFCC with CFCC, CFCCIF and CFCCIFS feature sets using *D1*, *D2* and *D3* feature vectors at various fusion factors α_f on the development set

Feature Set 1	Fusion Factor (α_f)											Feature Set 2
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
MFCC: D1	3.26	2.86	2.66	2.52	**2.43**	2.57	2.72	3.03	3.55	3.97	4.55	**CFCC: D1**
MFCC: D2	2.17	1.83	1.54	1.4	**1.32**	1.32	1.46	1.63	1.89	2.23	2.6	**CFCC: D2**
MFCC: D3	1.60	1.32	1.14	0.97	**0.89**	0.89	0.92	1.00	1.17	1.34	1.54	**CFCC: D3**
MFCC: D1	3.26	2.72	2.4	2.03	1.77	1.6	**1.52**	1.57	1.72	1.92	2.29	**CFCCIF: D1**
MFCC: D2	2.17	1.83	1.46	1.23	1.03	0.97	**0.89**	0.89	0.97	1.14	1.4	**CFCCIF: D2**
MFCC: D3	1.60	1.37	1.14	1.00	0.86	0.83	**0.83**	0.92	1.03	1.17	1.52	**CFCCIF: D3**
MFCC: D1	3.26	2.69	2.29	1.89	1.6	1.43	**1.37**	1.37	1.46	1.6	1.89	**CFCCIFS: D1**
MFCC: D2	2.17	1.74	1.37	1.09	0.89	0.8	**0.71**	0.74	0.77	0.92	1.06	**CFCCIFS: D2**
MFCC: D3	1.60	1.29	1.06	0.92	0.77	0.66	**0.66**	0.71	0.8	0.92	1.23	**CFCCIFS: D3**

Score-level fusion is carried as per eq. (3.6)

Dependency on spoofing algorithms: To check the discriminative property of the proposed feature set in terms of the dependency of the spoofing algorithm, the systems were trained on individual spoofs and tested on all the spoofs of the development set. The development set consists of SS and VCS spoofs which are further generated by different algorithms. Instead of considering the known and unknown attacks, we further break unknown attacks into two categories of 'same type' and 'different type'. However, the case of "same type" and "different type" of conditions is completely different than the "known" and "unknown" case mentioned for the ASV spoof database. For example, for training with S1 VC spoof: testing with speech from S1 algorithm itself is 'known', testing with speech from another VC-based algorithm (i.e., S2 and S5) is 'same type' and testing with speech from any SS-based spoofing algorithm (i.e., S3 and S4) is 'different type'. Average of the same type and different type constitutes 'unknown' attacks.

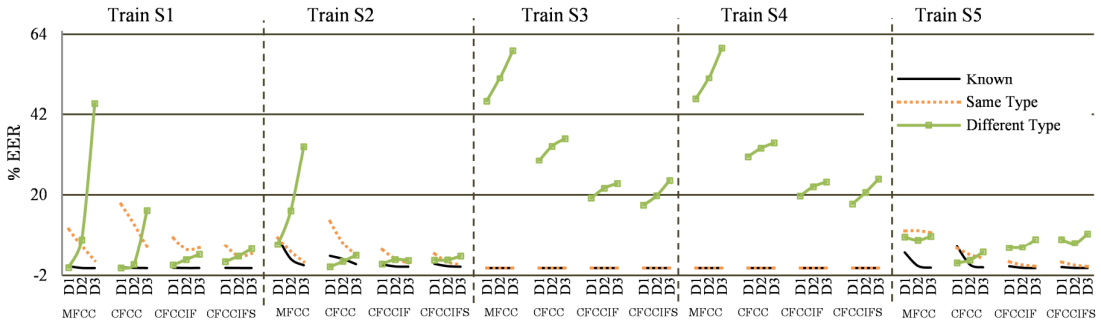


Figure 4.12: The % EER for known, same and different type of attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using different feature vectors (i.e., *D1*, *D2*, *D3)* and tested on the development dataset.

From Figure 4.12, it is observed that each of the features works well for known attacks (shown by black solid line). The SS spoofs *S3* and *S4* obtained *0.0* % EER when tested with itself. For VCS spoof, when tested by itself, the % EER decreased with MFCC, CFCC, CFCCIF and CFCCIFS feature sets with the *D3* feature vector. Next, for the 'same type' of attacks, SS spoofs (*S3* and *S4*) performed the best to detect each other. However, VC-based *S1* and *S2* that uses STRAIGHT vocoder identified VC-based *S5* spoof generated using MLSA vocoder with an average *10.7* % and *8.7* % EER and detected each other with *3.4* % (*S2*) and *0.12* % (*S1*) EER, respectively. On the other hand, *S5* spoof detected *S1* with *2.4* % and *S2* with *5.53* % EER. Thereafter, for 'different type' of spoof, VCS spoof detected synthetic speech to a certain extent, i.e., STRAIGHT-based *S1* and *S2* VCS method detected STRAIGHT-based *S3* and *S4* quite well with CFCCIF and CFCCIFS features. However, *S5* VCS spoof could not detect SS spoof well. Likewise, the SS spoof, when tested with VCS, gave very large % EER. For MFCC, around *50* % EER is observed which decreases to around *20* % for CFCCIFS features. The % EER increases for *D3* feature vector on testing with a different type of spoof, especially for MFCC. On the whole, the trend showed decrease in EER from MFCC to CFCCIFS features.

It was observed that, VCS spoof detected SS spoof to a certain extent. However, SS trained models could not detect VCS spoofs (this is also indicative in Table 4.4). Figure 4.13 shows the true (dotted) and false (solid) scores distribution of the testing data when trained only with *S1* VCS spoof (top panel) and *S3* SS spoof (bottom panel). The scores are shown for *D1* features vector of MFCC, CFCC, CFCCIF and CFCCIFS. While training with VCS spoof alone, features for synthetic speech could probably be captured and hence, the SS were detected in the testing phase which can

be observed from the single distribution of the false scores. Therefore, as models trained on VCS detected both SS and VCS, there is a single distribution of false scores. However, the models trained on SS could not detect VCS, resulting in an M-like distribution of false scores. The part of the M-shaped distribution that is overlapping with the true scores (dotted line) is likely due to scores from the VCS used in testing. The SS spoof model had been trained with features specific to it and hence, could not detect VCS resulting in large overlapping regions leading to increased % EER. The SS spoof model had been trained with features specific to it and hence, could not detect VCS resulting in large overlapping regions leading to increased % EER. Amongst all the feature sets considered here, CFCCIFS had the least overlap between true and false score distribution resulting in low % EER and better performance.

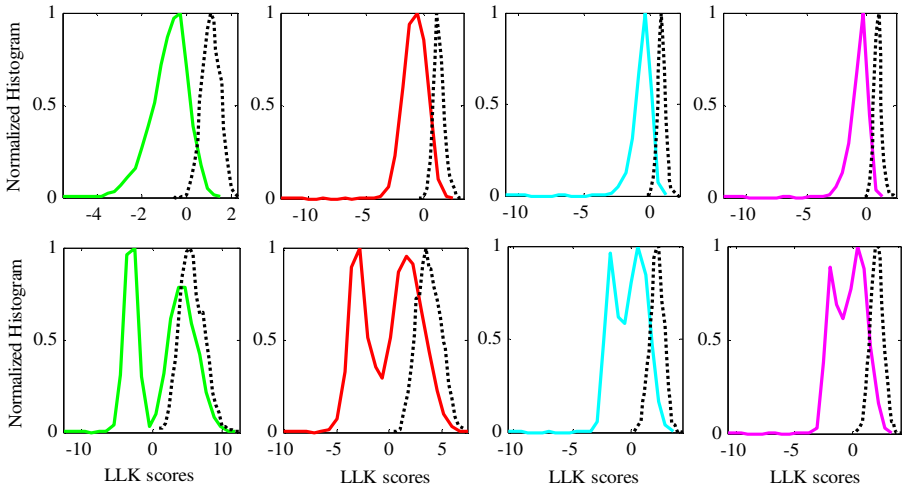


Figure 4.13: The distribution (i.e., normalized histogram) for true scores (dotted line) and false scores (solid line) for *12-D* static features extracted from MFCC (green), CFCC (red), CFCCIF (cyan) and CFCCIFS (magenta) when trained with *S1* VCS spoof (top panel) and *S3* SS spoof (bottom panel).

Table 4.4 shows % EER of known and unknown attacks when trained with individual spoofs and tested on the development set. The VCS spoofs gave less % EER for unknown attacks as it detected SS spoof to a certain extent. On the other hand, SS spoof gave very high % EER, i.e., > *30 %* with MFCC and > *10 %* with CFCCIFS. The shaded cells show best performance of known (Kn) and unknown attacks (Ukn) attacks.

Table 4.4: EER (in %) for known and unknown attacks when trained on individual spoofs and tested on the development set

		Feature sets							
		MFCC		CFCC		CFCCIF		CFCCIFS	
Train	Dim.	Kn	Ukn	Kn	Ukn	Kn	Ukn	Kn	Ukn
S1 (VC)	D1	0.40	5.48	0.17	8.8	0.08	4.62	0.05	3.97
	D2	0.03	6.96	0.06	6.57	0.02	3.71	0.02	3.25
	D3	0.00	23.48	0.01	10.83	0.03	4.63	0.01	4.59
S2 (VC)	D1	8.39	7.45	3.31	6.63	0.98	3.08	1.14	3.04
	D2	2.49	10.11	2.45	4.34	0.43	2.41	0.51	1.90
	D3	0.76	17.51	1.06	3.56	0.36	1.74	0.36	2.10
S3 (SS)	D1	0.00	34.21	0.00	22.06	0.00	14.35	0.00	12.93
	D2	0.00	38.93	0.00	25.06	0.00	16.37	0.00	14.83
	D3	0.00	44.65	0.00	26.61	0.00	17.32	0.00	17.94
S4 (SS)	D1	0.00	34.65	0.00	22.95	0.00	14.8	0.00	13.19
	D2	0.00	38.97	0.00	24.66	0.00	16.69	0.00	15.51
	D3	0.00	45.19	0.00	25.74	0.00	17.62	0.00	18.23
S5 (VC)	D1	4.22	9.46	5.91	3.43	0.49	3.66	0.28	4.79
	D2	0.60	9.06	0.89	2.89	0.08	3.30	0.04	3.83
	D3	0.12	9.34	0.21	3.56	0.01	4.19	0.03	4.92

Kn=known, Ukn=Unknown

4.5.3.3 Results on the Evaluation Set of ASVspoof challenge Database

On the development set, it was observed that instead of using static or dynamic features alone, their combination (i.e., *D3* feature vector) gives less % EER. In addition, score-level fusion of MFCC and cochlear filter-based features one at a time, i.e., MFCC with CFCC (or CFCCIF or CFCCIFS) features gave the least % EER.

Table 4.5: EER (in %) for score-level fusion of MFCC with CFCC, CFCCIF and CFCCIFS feature sets using *D1*, *D2* and *D3* feature vectors at various fusion factors α_f on the evaluation dataset

Feature Set 1	Fusion Factor (α_f)											Feature Set 2
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
MFCC: D1	5.5	4.86	4.49	4.21	4.09	4.03	4.02	4.15	4.33	4.64	4.98	CFCC: D1
MFCC: D2	5.49	4.23	3.67	3.19	2.81	2.58	2.45	2.4	2.45	2.58	2.78	CFCC: D2
MFCC: D3	4.26	3.19	2.67	2.26	1.99	1.79	1.69	1.62	**1.61**	1.63	1.74	CFCC: D3
MFCC: D1	5.5	4.92	4.46	4.04	3.74	3.51	3.35	3.23	3.16	3.19	3.37	CFCCIF: D1
MFCC: D2	5.49	4.52	3.96	3.53	3.13	2.79	2.55	2.38	2.3	2.28	2.34	CFCCIF: D2
MFCC: D3	4.26	3.46	2.99	2.66	2.39	2.19	2.03	1.91	**1.87**	1.91	2.07	CFCCIF: D3
MFCC: D1	5.5	4.84	4.31	3.84	3.46	3.17	2.98	2.85	2.74	2.74	2.81	CFCCIFS: D1
MFCC: D2	5.49	4.41	3.74	3.24	2.81	2.45	2.13	1.91	1.8	1.75	1.81	CFCCIFS: D2
MFCC: D3	4.26	3.36	2.79	2.41	2.06	1.8	1.6	1.49	**1.45**	1.48	1.6	CFCCIFS: D3

Score-level fusion is carried as per eq. (3.6)

Results of score-level fusion: Table 4.5 shows the results in % EER on the evaluation data after fusion of MFCC features with CFCC (or CFCCIF or CFCCIFS) sets using *D1*, *D2* and *D3* feature vectors. Table 4.5 shows that score-level fusion of

D3 feature vector gave best results for all the features. Unlike the case of development set, where the best % EER was obtained at α_f *=0.4* for CFCC and α_f *=0.6* for CFCCIF and CFCCIFS, for the evaluation set, the fusion factor changes to α_f *=0.8*, due to the presence of unknown attacks. With α_f *=0.8,* the cochlear-based features contribute more in reducing the % EER in the case of unknown attacks. In fact, the *1.6* % EER of the *D3* feature vector for CFCCIFS alone is almost equal to the least EER of *1.45 %* after fusion.

Results on individual attacks: The attack-dependent % EER for the individual spoofs of the evaluation set using all the four feature vectors (i.e., MFCC, CFCC, CFCCIF, and CFCCIFS) are shown in Table 4.6.

Table 4.6: EER (in %) in terms of individual attacks, average known attacks, average unknown attacks, average with and without *S10* spoof for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using all the feature vectors and for the score-level fusion of MFCC with CFCC, CFCCIF and CFCCIFS feature sets (using *D3* feature vector) at selected α_f on the evaluation dataset

Feature Vectors	Feature Sets	Individual Spoofs										Average			
		Known					Unknown								
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn	Ukn	w/o S10	Avg.
D1 (s)	MFCC	0.1	3.8	0.0	0.0	4.5	3.6	0.4	0.0	1.0	41.5	1.7	9.3	1.48	5.5
	CFCC	0.1	2.6	0.0	0.0	11.3	4.8	0.4	0.1	1.2	29.2	2.8	7.1	2.27	5.0
	CFCCIF	0.0	0.5	0.0	0.0	3.3	1.3	0.1	0.2	0.2	28.2	0.8	6.0	0.62	3.4
	CFCCIFS	0.0	0.5	0.0	0.0	2.6	1.0	0.1	0.2	0.2	23.5	0.6	5.0	0.51	2.8
D2 (s+Δ)	MFCC	0.0	1.6	0.0	0.0	1.6	1.8	0.1	0.0	0.0	49.6	0.7	10.3	0.56	5.5
	CFCC	0.1	2.5	0.0	0.0	5.0	2.5	0.3	0.1	0.6	16.8	1.5	4.0	1.23	2.8
	CFCCIF	0.0	0.2	0.0	0.0	1.5	0.7	0.1	0.4	0.1	20.5	0.4	4.3	0.33	2.3
	CFCCIFS	0.0	0.2	0.0	0.0	1.3	0.5	0.1	0.5	0.1	15.4	0.3	3.3	0.3	1.8
D3 (s+Δ+ΔΔ)	MFCC	0.0	1.0	0.0	0.0	0.8	0.9	0.1	0.0	0.0	39.7	0.4	8.2	0.31	4.3
	CFCC	0.0	1.4	0.0	0.0	2.3	1.0	0.1	0.1	0.2	12.3	0.8	2.7	0.57	1.7
	CFCCIF	0.0	0.7	0.0	0.0	2.2	1.0	0.2	0.9	0.3	15.4	0.6	3.6	0.59	2.1
	CFCCIFS	0.0	0.5	0.0	0.0	1.7	0.7	0.1	1.0	0.2	11.7	0.5	2.7	0.47	1.6
D3 α_f=0.8	MFCC+CFCC	0.0	0.7	0.0	0.0	1.3	0.7	0.1	0.0	0.1	13.1	0.4	2.8	0.33	1.6
	MFCC+CFCCIF	0.0	0.4	0.0	0.0	1.0	0.5	0.0	0.1	0.0	16.7	0.3	3.5	0.22	1.9
	MFCC+CFCCIFS	0.0	0.2	0.0	0.0	0.7	0.3	0.0	0.1	0.0	13.0	**0.2**	**2.7**	**0.16**	**1.4**

Score-level fusion is carried as per eq. (3.6), Kn=known, Ukn= Unknown, w/o S10=Average without S10, Avg. = Average of S1-S10

It is observed that when trained with VCS and SS jointly, the known and unknown attacks are detected quite well (except *S10*). For *D3* feature vector, MFCC features achieved least *0.37 %* EER for known attacks. However, it achieved very high *40 %* EER for *S10* which increased the average EER to *4.26 %*. The CFCCIFS feature set gave *0.45 %* for known attacks and lowest EER of *2.73 %* on unknown attacks with *11.7 %* EER for *S10* spoof. The CFCC feature set also obtained less % EER for *S10* spoof. However, its % EER for other spoofs was more than that of CFCCIF and CFCCIFS. Amongst known attacks, *S2* and *S5* spoof were difficult to detect and for

unknown attacks, the vocoder-independent *S10* spoof was toughest, followed by *S6*, *S9*, *S8,* and *S7*. Overall, CFCCIFS feature set works quite well for known attacks and detected unknown spoofed speech even if a similar type was not available during training.

Figure 4.14 shows the % EER for known and unknown attacks with *D3* feature vector at various fusion factors when MFCC is fused with either CFCC or CFCCIF or CFCCIFS. While MFCC features, when used alone, gave least % EER for known attacks, the cochlear filter-based features gave least % EER for unknown attacks even without fusion with MFCC. Thus, some contribution of MFCC is required for best performance with known attacks. Table 4.6 shows the % EER for optimum α_f of *0.8* for cochlear-based features. Amongst all score-level fusions, MFCC+CFCCIFS combination performs relatively best.

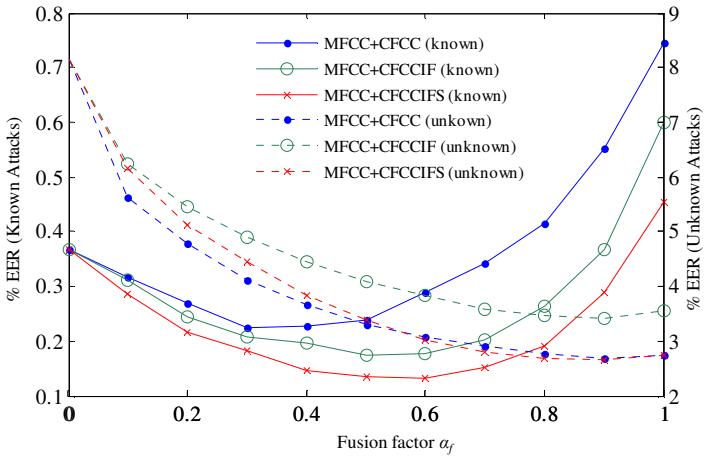


Figure 4.14: The % EER of known attacks (solid line) and unknown attacks (dashed line) for *D3* feature vector on fusion of MFCC with CFCC (blue), CFCCIF (green) and CFCCIFS (red).

Discussion on the DET curves: The DET curves for MFCC, CFCC, CFCCIF and CFCCIFS feature sets, when used alone, are shown in Figure 4.15 (a). It is observed that the FRR of MFCC was very high for a given FAR which is not suitable for ASV systems. From MFCC to CFCC, there is a significant decrease in FRR (as shown in Figure 4.15 (a) by dotted region) which further reduces for CFCCIFS. Figure 4.15 (b) shows DET curves for CFCC, CFCCIF and CFCCIFS after fusion with MFCC for α_f =*0.8*. A clear improvement in both FRR and FAR is observed with MFCC+CFCCIFS than with CFCC and CFCCIF. The results of score-level fusion of MFCC+CFCCIF with α_f =*0.6* was submitted at the ASVspoof 2015 challenge (as decided from the

results on development set). However, here MFCC+CFCCIFS gives better % EER and the best performance amongst all the other fusion combinations.

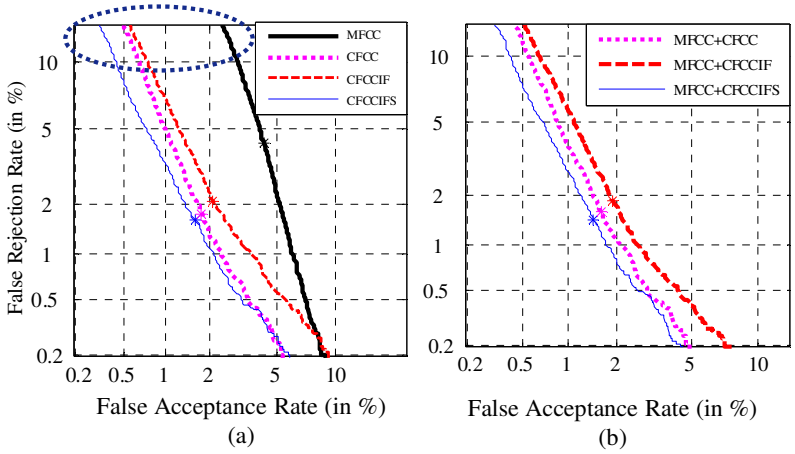


Figure 4.15: DET curves on the evaluation set for (a) MFCC, CFCC, CFCCIF and CFCCIFS feature sets used alone (b) score-level fusion of MFCC and CFCC, MFCC and CFCCIF and MFCC and CFCCIFS with α_f =0.8.

Dependency on spoofing algorithms: Just as in the case of development set, the testing of the entire evaluation data is carried out when trained on individual *S1*, *S2*, *S3*, *S4* and *S5* spoofs. For the development set, known and same types of spoofs were easily identified. However, for 'different type' of spoof on training with VCS spoofs, the SS spoofs were identified while the reverse case was not true. The same analysis is continued here using 'same type' and 'different type' of spoofed speech. The *S10* spoof is considered separately as it is non-vocoder type and its performance highly affects the average % EER. The interpretations for 'known type' remains similar as discussed for Figure 4.12 (and hence, not shown here again).

Figure 4.16 shows that as in the development set, the trend is similar, i.e., for the 'same type' of spoofs, SS identified its same type with almost *0.00* % EER. The VCS spoof identified its same type quite well and the % EER decreased from MFCC to CFCCIFS features. On the other hand, for 'different type' of spoof, VCS spoofs gave less % EER when tested on SS attacks as compared to SS spoof that gave very high % EER on testing with VCS spoof. Overall, on training with *S1*, *S2* and *S5* (VCS spoof), MFCC feature set gave an EER of *4.71 %, 4.57 %* and *2.12 %* and CFCCIFS gave an average EER of *1.43 %, 1.00 %* and *2.10 %* for all vocoder-based (*S1-S9*) spoofs averaged over all dimensions. This analysis was independent of *S10* spoof.

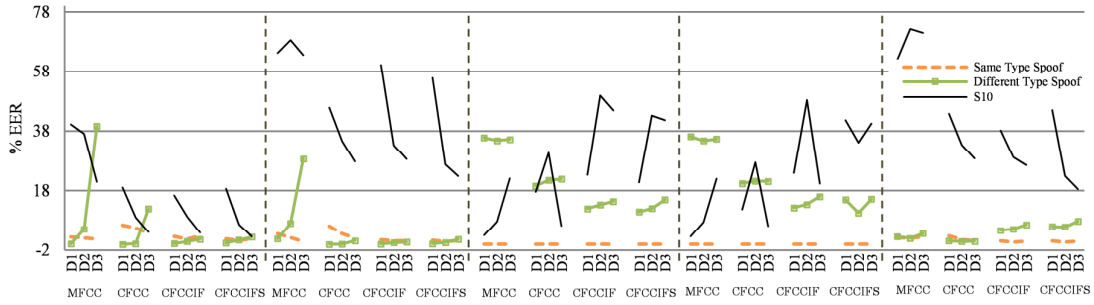


Figure 4.16: The % EER for same type, different type and *S10* attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using different vectors (i.e., *D1*, *D2*, *D3*) and tested on the evaluation dataset.

Considering *S10* separately, VCS spoofs (*S1*, *S2* and *S5*) when tested with *S10,* gave large % EER for MFCC feature set, i.e., the detection rate were around *20-70 %* with MFCC. The % EER gradually decreased to *15-25 %* when CFCCIFS features are used with *D3* feature vector. Interestingly, when trained on *S1* spoof with CFCCIFS using *D3* feature vector, as low as *2.6 %* EER is achieved. Similarly, on testing with SS trained models, *S10* gave as low as *3 %* with MFCC using *D1* feature set which increased to *10-50 %* when other features were used. For MFCC, the % EER increases from *D1* to *D3* feature vector, while the pattern of EER for cochlear filter-based features is found to be random. On listening to MARY TTS speech utterances (*S10* spoof) from SAS database [**19**], they were found to be *unintelligible* [**123**]. Thus, it should be possible to identify this kind of spoofed speech. However, the modeling needs to be done appropriately and techniques to deal with vocoder-independent speech needs to be explored further. On the whole, the proposed CFCCIFS feature set performed better due to the use of auditory filterbank than triangular filterbank and due to the notion that human speech production system produces speech in continuum manner rather in a frame-by-frame pattern which when embedded in CFCCIFS gives better spoof detection results.

4.5.3.4 Results on the Blizzard Challenge 2012 Database

The Blizzard Challenge 2012 database consists of unknown spoofing algorithms of both HTS and USS system. In this case, the systems are trained on the ASV spoof database and tested on the Blizzard Challenge 2012 database. The % EER was obtained using *A* (*100* utterances of natural speech signal) and one spoofed system from *B-K* (*100* utterances each). The results of detection are shown in Table 4.7.

As observed from Table 4.7, the % EER by using cochlear-based features are much better than MFCC. The system *B* was the benchmark system and was found to be toughest to detect with minimum EER obtained of *45* % with CFCC features. It is to be noted that for the Blizzard Challenge 2012 evaluation, the CFCCIF/CFCCIFS features did not always perform best as in the case of ASV spoof database. As in Figure 3.6, systems *D* (hybrid), *E* (statistical), *J* (diphone) and *K* (statistical) had an MOS score of ≤ *2*. The minimum % EER obtained for these attacks are *23*, *2, 41* and *6*. A low % EER was observed for statistical-based spoofing attacks. The systems *C* (hybrid), *F* (USS) and *I* (USS) had an MOS >*3* and were detected with an EER of *26* %, *9* % and *31* %, respectively. Thus, there was no exact correlation between the % EER of the features and the MOS scores for the Blizzard dataset.

Table 4.7: EER (in %) for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using *D1*, *D2* and *D3* feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2012 database

Blizzard 2012	Systems	Feature Sets											
		MFCC			CFCC			CFCCIF			CFCCIFS		
		D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3
USS	**B**	98	77	67	48	47	**45**	51	47	58	53	53	50
Hybrid	**C**	40	46	47	37	28	26	33	31	31	36	33	**23**
Hybrid	**D***	65	66	42	36	42	47	29	35	41	**23**	34	36
HMM	**E***	44	82	61	12	6	11	3	3	11	**2**	2	8
USS	**F**	22	24	15	22	17	20	**9**	12	30	11	14	13
USS	**G**	8	29	27	11	7	11	**5**	6	25	**5**	6	10
HMM	**H**	12	38	3	1	1	3	3	3	2	**1**	2	2
USS	**I**	98	97	69	34	31	**31**	46	48	49	42	45	45
Diphone	**J***	64	69	69	45	41	**41**	58	53	59	61	59	53
HMM	**K***	92	67	73	30	33	42	13	**6**	37	7	12	20

* systems with lower MOS from *1* ≤ *2*

4.5.3.5 Results on the Blizzard Challenge 2014 Database

The results for testing on Blizzard Challenge 2014 database for Gujarati and Hindi languages are shown in Table 4.8 and Table 4.9, respectively. For the Gujarati language, the least % EER was obtained using MFCC features with *D3* feature set. The % EER was very high for all cochlear-based features. For MFCC, the HMM-based systems gave less % EER (as training is done on vocoder-based spoofs) and USS-based *G* system gave a high EER of *34* %. None of the cochlear-based features could perform better than MFCC features.

Table 4.8: EER (in %) for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using *D1*, *D2* and *D3* feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2014 database for the Gujarati language

Blizzard 2014	Gujarati Systems	Feature Sets											
		MFCC			CFCC			CFCCIF			CFCCIFS		
		D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3
HMM	**C**	37	13	**4**	38	35	35	87	90	65	82	82	59
HMM	**D**	67	6	**1**	18	18	14	94	90	65	88	88	39
HMM	**E**	13	21	**4**	82	51	44	100	90	100	99	99	71
HMM-DNN	**F**	75	23	**6**	97	73	73	99	90	99	98	98	78
USS	**G**	67	48	**34**	84	85	84	100	98	89	99	99	87
HMM	**H**	55	24	**24**	86	58	62	100	87	87	100	95	54

* wavefiles for baseline system B and system I are not available

Table 4.9: EER (in %) for MFCC, CFCC, CFCCIF and CFCCIFS feature sets using *D1*, *D2* and *D3* feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2014 database for the Hindi language

Blizzard 2014	Hindi Systems	Feature Sets											
		MFCC			CFCC			CFCCIF			CFCCIFS		
		D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3
HMM	**B***	19	**5**	14	87	68	49	80	63	79	84	82	63
HMM	**C**	7	**4**	6	20	14	10	12	4	16	14	12	12
Hybrid	**D**	39	**25**	62	77	70	67	70	62	70	76	70	63
HMM	**E**	6	**5**	7	72	40	15	75	43	86	73	65	27
HMM-DNN	**F**	32	**7**	19	75	60	46	72	52	55	82	68	40
USS	**G**	51	52	52	34	26	28	27	26	24	31	27	**24**
HMM	**H***	10	11	30	42	13	2	15	**0**	38	14	2	**0**
HMM	**K**	72	8	32	38	15	4	20	6	16	29	22	**2**

* systems with lower MOS from *1* ≤ *2* (wavefiles for system I are not available)

For Hindi, similar to Blizzard Challenge 2012 dataset, no consistency was observed with respect to a definite feature giving better results for all the systems considered here. However, it was observed that HMM systems were definitely detected with EER around < *11* %. The hybrid system *D* achieved the least EER of *25* % with MFCC and USS-based system *G* achieved an EER of *24* % with CFCCIFS features. It can be observed from the analysis of Blizzard Challenge data both in English and other languages that there is an inconsistency of features in detecting spoofing attacks. On the whole, HMM-based systems were easily detected than the USS synthesis systems. However, this observation is not uniform as in the case of Blizzard Challenge 2012 database, the USS-based systems *F* and *G* were detected with lower % EER. Hence, there is a need for generalized countermeasures.

4.6 Subband Autoencoder (SBAE)

4.6.1 Introduction to Autoencoder (AE)

The Autoencoder (AE) is such a network which uses Deep Neural Network (DNN) or Restricted Boltzmann Machine (RBM) to extract low-dimensional information from high-dimensional raw data [**173**]. The AEs have been used in various applications such as de-noising front-end for Automatic Speech Recognition (ASR) task, in finding a mapping between noisy and clean speech spectrum for noise reduction in ASR system, speech enhancement task and speech coding. Very recently, the AEs have been used for noise reduction in SV system. The deep AE has also been used for noise aware training for noisy ASR. Features learned by deep AE were also used for SPSS using DNN. Figure 4.17 shows the basic architecture of AE.

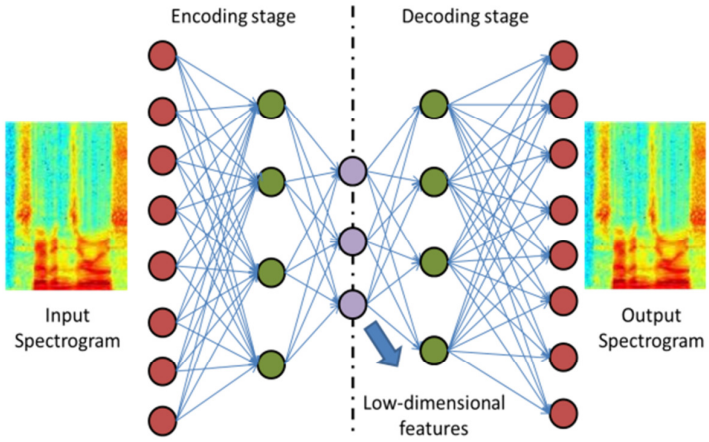


Figure 4.17: Architecture of the Autoencoder (AE). Adapted from [173**].**

Despite the various properties, AE features are not popular as acoustic features in most of the speech technology applications. The inability to control the form of the representation which is learned by AE leads some researchers to criticize them as uninterruptable black boxes [**174**]. To overcome this limitation, many variants of the AE have been proposed. A new architecture called transforming AE was used to detect acoustic events in speech signal for the ASR task. Phone recognition task was done using mean-covariance RBM [**174**]. In [**175**], authors proposed an architecture of AE in which decoding block was constrained for stretching and compressing frequency-domain for ASR task.

4.6.2 Subband Autoencoder (SBAE)

In this thesis, a modified architecture of AE, i.e., Subband Autoencoder (SBAE) is used for feature extraction from the speech spectrum. Proposed architecture uses domain-specific knowledge about speech processing and incorporates it in the architecture of AE. Inspired by Human Auditory System (HAS), speech is generally processed in subbands. Therefore, in SBAE, the connectivity of units in traditional AE is restricted in such a way that each unit in first hidden layer captures the information of a particular band of the speech spectrum. This property of our architecture makes it more suitable for several speech technology applications. The features extracted by SBAE are used for the present task of spoof detection. The main difference between proposed SBAE architecture and existing architecture of AE is the connectivity of neurons or units immediately after the input layer [**173**]. This architecture of the SBAE is shown in Figure 4.18.

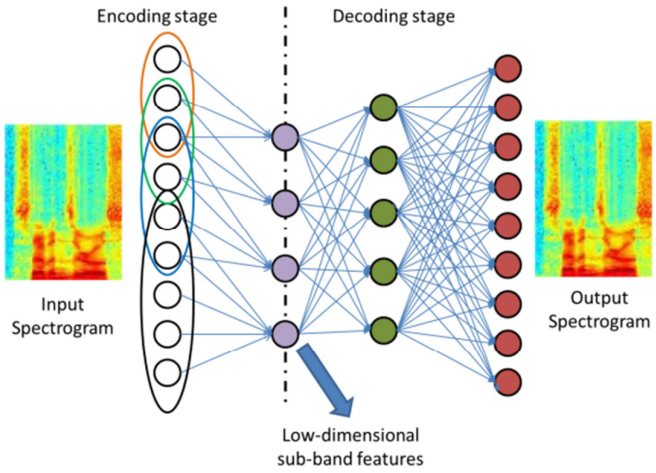


Figure 4.18: Architecture of the Subband Autoencoder (SBAE). After [89**].**

In AE, each unit in the layer immediately after input layer is connected with all the units of the previous layer. While in the case of proposed SBAE, the connectivity is restricted. In the proposed SBAE architecture, each unit of the first hidden layer is connected with a particular frequency band of input spectrogram. Hence, each unit in the first layer will encode the information about that particular frequency band with which it is connected. The decoding structure is same as that of traditional AE with full connectivity [**173**]. The band structure of restricted connectivity for neurons is same as Mel filterbank, implying one neuron in the first layer is connected with

the frequencies of one Mel filterbank. This architecture closely resembles to HAS and provides more meaningful information than AE in the case of the speech signal. Mathematically, operation of the subband layer can be represented as follows:

$$a_j = f(\sum_i W_{ij}^1 \times x_j), \tag{4.18}$$

where a_j is j^{th} subband feature, x_j is short-time power corresponding to j^{th} filterbank and W_{ij} are weights corresponding to j^{th} subband feature, f is the nonlinear activation function of the neuron. The functionality of preceding layers of SBAE is same as of traditional AE [**173**]. The input to the SBAE is the linearly scaled spectrogram and the input is log-compressed. Proposed SBAE architecture can be trained by back-propagation similarly as an AE. The learnt a_j can be used as low-dimensional features for several other speech technology applications.

4.6.3 Analysis of SBAE on Spoofed Speech

To observe the effect of SBAE features for the SSD task, the SBAE representation is observed for natural and various types of spoofed speech as shown in Figure 4.19.

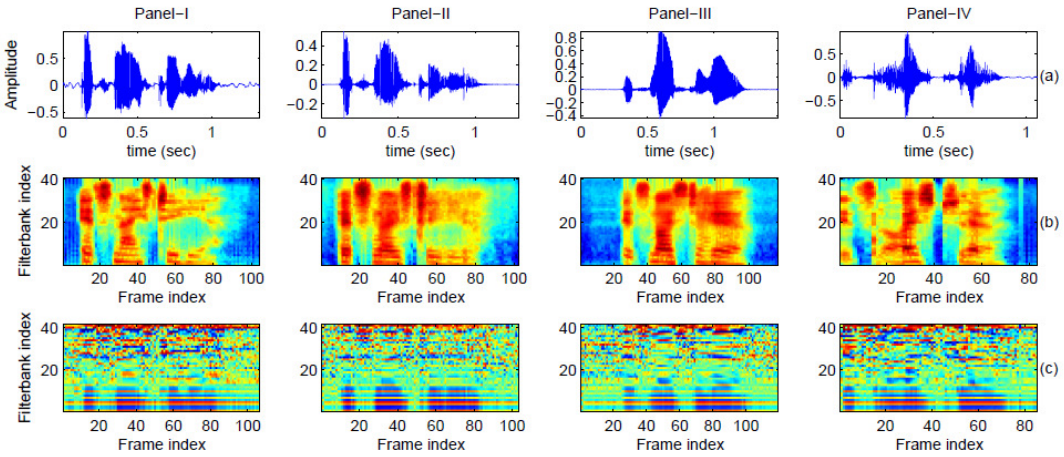


Figure 4.19: (a) Speech signal waveform, (b) Mel filterbank energies and (c) SBAE features energies for Panel I: natural speech, Panel II: vocoder-based VCS, Panel III: vocoder-based SS and Panel IV: USS-based MARY TTS.

It can be observed that both MFCC and SBAE features show variations for natural and different types of spoofs. Hence, both of these features can be used for spoof detection task. Moreover, both feature sets are invertible, implying speech spectrum can be reconstructed using both features (while it may not be strictly necessary for classification problem). To quantify reconstruction ability of both the features,

average Log Spectral Distortion (LSD) between original spectrum $P(\omega)$ and reconstructed spectrum $\hat{P}(\omega)$ was calculated as follows:

$$LSD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]^2 d\omega} \ . \tag{4.19}$$

For *50* natural utterances of ASVspoof 2015 database, the LSD in the case of proposed SBAE features was *5.01 dB* and using Mel filterbank energies it was *9.04 dB*. As observed in Figure 4.19, the proposed features do not show many variations in low-frequency regions for natural speech and also for different conditions of spoofing attacks. It is also observed that the proposed features are more sensitive to small variations in the spectrum due to nonlinear processing. This effect can be seen by observing features of two consecutive frames in Figure 4.19. Unlike Mel filterbank energies, proposed features vary more for consecutive frames (in time-domain). Thus, SBAE features may capture more dynamic information of speech spectrum. Similar findings for AE features were observed in [**176**].

The most important property of SBAE feature for spoof detection can be seen in Figure 4.20 (a) and Figure 4.20 (c) which shows variance of SBAE features and the Mel filterbank energies for natural utterances and utterance synthesized using USS (*S10* system in ASV spoof challenge database). The USS-based spoof is proven to be the most difficult to detect amongst all the synthesis techniques. The reason behind this is that USS system uses different units of natural speech and concatenates the units to generate output speech signal according to text input. Since USS systems use segments of natural utterance, they sound very natural (though not always intelligible) and difficult to discriminate from natural utterances. Due to this reason, state-of-the-art features such as MFCC, which works very well on another kind of attacks such as VCS and HMM-based SS, gives poor results on synthetic speech generated by USS systems [9]. This effect is visible in Figure 4.20 (a) and Figure 4.20 (c). The variance of higher-order Mel filterbank energies is almost similar in the case of natural speech and USS-based speech. However, the Mel filterbank energies show different variance for another kind of speech such as VCS (Figure 4.20 (b)) and SS (Figure 4.20 (d)). On the other hand, higher order SBAE features show different variance for all the types of spoofed speech. The difference between the variance of natural speech and speech synthesized by USS is clearly visible. Hence, SBAE

features may work better than MFCCs in the case of USS speech. Moreover, due to the very high and low variance of the SBAE features for different types of spoof, by including their dynamic variations as countermeasure, the performance is likely to be improved than using only static information.

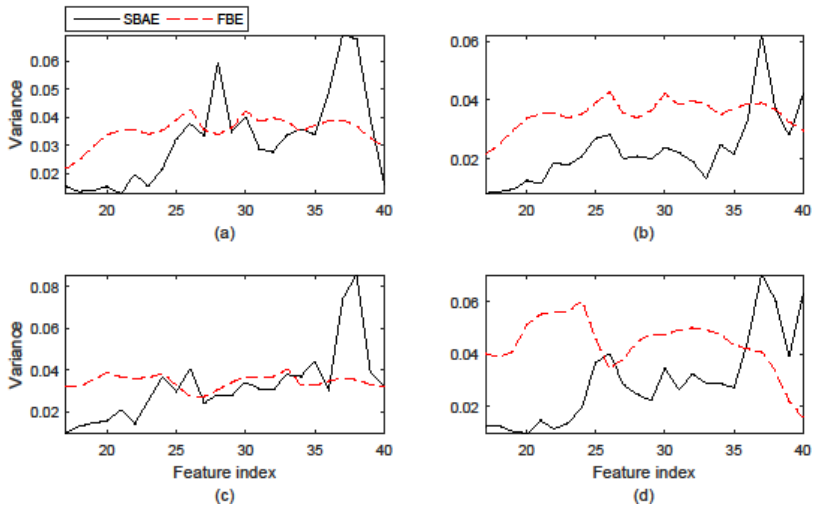


Figure 4.20: Variance of higher-order Mel filterbank energies (FBEs) and SBAE features for (a) natural speech, (b) vocoder-based VCS, (c) USS-based MARY TTS speech and (d) vocoder-based SS.

4.6.4 Experimental Results

4.6.4.1 Parameterization for SBAE features

For feature extraction of SBAE, the speech signals were divided into frames with *25* ms duration and *50* % overlap. The STRAIGHT spectrum was used for feature extraction using SBAE [**132**]. The configuration of the network was *513-40-250-513*, implying *513* units in the input layer, *40* units in subband layer, *250* units in the second layer and *513* units in the output layer. Input and output data were normalized between *0-1* for training. The SBAE trained on training data was used for feature extraction from validation and evaluation datasets. Here, *40* units in subband layer give *40* subband features. To compare the performance of proposed features with the *12-D* MFCC and with *12-D* cochlear filter features, the *40-D* SBAE features were converted to *12-D* features by the following process. As it can be observed from Figure 4.19, not all *40* SBAE features vary significantly for different types of speech. The SBAE features corresponding to lower bands have almost constant values for the natural and spoofed speech. The SBAE features for first *16*

subbands were removed and features corresponding to rest of the *24* subbands were used. Hence, SBAE features corresponding to higher subbands are considered for discrimination task. For further dimensionality reduction, the average value of two consecutive subband features was taken and *24* subbands were converted to *12* subbands. Hence, by this method, *12-D* feature vector was generated for comparison. As a similarity check, our preliminary experiments suggested that EERs on development set using *40-D* features and reduced *12-D* features were almost similar.

4.6.4.2 Results on the Development Set of ASVspoof challenge Database

The results on the development set for MFCC and SBAE are shown in Table 4.10. It is observed that on using static feature vector, the SBAE features gave an EER of *5.38 %* which is more than MFCC, CFCC, CFCCIF and CFCCIFS feature sets. On using the Δ features with the static features, for SBAE the EER is almost similar to MFCC and CFCC, i.e., *2.37* %, *2.17* %, and *2.60* %, respectively. Further with the use of ΔΔ features, the EER of SBAE reduces to *1.49* % which is better than MFCC, CFCC and CFCCIF features alone. Thus, dynamic features capture more spectral variation than only static features which reduces the % EER for SBAE features.

Table 4.10: EER (in %) for score-level fusion of SBAE with MFCC, CFCC, CFCCIF and CFCCIFS feature sets using *D1*, *D2* and *D3* feature vectors at various fusion factors α_f on the development set

Feature Set 1	Fusion Factor (α_f)											Feature Set 2
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
SBAE: D1	5.38	4.38	3.69	3.23	2.89	2.72	2.60	2.72	2.86	3.06	3.26	**MFCC: D1**
SBAE: D2	2.37	1.86	1.54	1.37	1.37	1.40	1.46	1.57	1.72	1.92	2.17	**MFCC: D2**
SBAE: D3	1.49	1.06	0.83	**0.71**	**0.71**	0.77	0.86	1.00	1.14	1.34	1.60	**MFCC: D3**
SBAE: D1	5.38	4.46	3.83	3.43	3.17	3.20	3.35	3.60	3.95	4.23	4.55	**CFCC: D1**
SBAE: D2	2.37	1.86	1.52	1.40	1.37	1.46	1.63	1.77	2.03	2.34	2.60	**CFCC: D2**
SBAE: D3	1.49	1.09	0.86	**0.77**	0.83	0.86	0.97	1.03	1.20	1.37	1.54	**CFCC: D3**
SBAE: D1	5.38	4.46	3.66	2.86	2.32	2.09	1.97	1.92	2.03	2.09	2.29	**CFCCIF: D1**
SBAE: D2	2.37	1.77	1.34	1.12	0.97	1.00	1.00	1.06	1.14	1.23	1.40	**CFCCIF: D2**
SBAE: D3	1.49	1.12	0.94	0.77	**0.77**	0.83	0.94	1.06	1.14	1.29	1.52	**CFCCIF: D3**
SBAE: D1	5.38	4.35	3.46	2.55	2.03	1.74	1.69	1.69	1.74	1.77	1.89	**CFCCIFS: D1**
SBAE: D2	2.37	1.72	1.29	1.00	0.80	0.80	0.86	0.86	0.92	0.94	1.06	**CFCCIFS: D2**
SBAE: D3	1.49	1.06	0.86	0.69	**0.63**	0.69	0.74	0.80	0.89	1.03	1.23	**CFCCIFS: D3**

Score-level fusion is carried as per eq. (3.6)

Results of score-level fusion: The results of score-level fusion of SBAE features with other system-based features are shown in Table 4.10. On the other hand, when traditional *36-D* AE features were used an EER of *7.9* % was obtained. Hence, in this work, we do not consider AE features for further analysis. It is observed that with a

fusion factor of around α_f = *0.3* or *0.4* the EER achieved was *0.71* % which is almost half as compared to using MFCC and SBAE features alone. Thus, SBAE features captured complementary information as compared to MFCC features. Similar results were observed for CFCC and CFCCIF features as well. For CFCCIFS features, the lowest EER of *0.63* % is obtained on score-level fusion with SBAE features. However, this decrease is due to efficient spoof detection by CFCCIFS features. Therefore, from the development set a fusion factor of α_f = *0.3* for MFCC α_f = *0.4* for the cochlear-based features can be considered optimum for score-level fusion.

Dependency on spoofing algorithms: As discussed in Section 4.5.3.2, to evaluate the spoof dependency, the SBAE features are evaluated for known type, same type and different type of attacks. As shown in Figure 4.21, similar to the cochlear-based features, the % EER decreases for known type of attacks with the dynamic features.

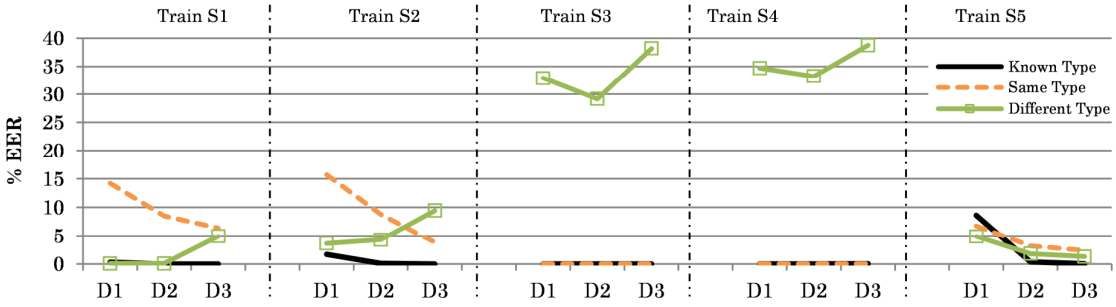


Figure 4.21: The % EER for known, same and different type of attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for SBAE feature set using different feature vectors (i.e., *D1*, *D2*, *D3)* and tested on the development dataset.

For the same type of attacks, the SS spoof identifies SS with *0.00* % EER. For VCS, the % EER decreases with the use of dynamic features. The *S5* spoof identifies *S1* and *S2* much better than *S1* and *S2* could identify *S5*. The *S5* spoof uses MLSA vocoder and *S1* and *S2* use STARIGHT vocoder. For the different type of spoof, it was observed that as in the case of cochlear-based features, the VCS could identify SS, however, the SS could not identify VCS spoof. In addition, it is observed that for different type of attacks, the % EER increases with the dynamic information of features. This was also observed in MFCC and cochlear-based feature sets. Therefore, for system-based features, the dynamic information in the feature vector is not able to capture information about different type of attacks.

4.6.4.3 Results on the Evaluation Set of ASVspoof challenge Database

The results of the development set were based on known attacks, this is because the development set has same type of spoof as used in the training. However, the anti-spoofing ability of the countermeasure depends on the performance in presence of unknown spoofing attacks. To that effect, the performance of the features is tested on the evaluation set which consists of unknown vocoder-based spoof and one vocoder-independent spoof. The results in % EER for SBAE features are shown in Table 4.11. It is observed that the presence of dynamic features improves the performance significantly for SBAE features from an average *9.08* % EER with static features to *2.49* % with the *D3* feature vector. Considering the results obtained in Section 4.5.3.3, for known attacks (*S1-S5*) the EER of MFCC, CFCC, CFCCIF, CFCCIFS and SBAE features are *0.37* %, *0.8* %, *0.6* %, *0.5*, and *1.06* %, respectively, using *D3* feature vector. That is, among all the features MFCC works best for known attacks. Likewise for the unknown attacks, the performance was *10.3* %, *4.0* %, *4.3* %, *3.3* % and *3.92* % for MFCC, CFCC, CFCCIF, CFCCIFS and SBAE features, respectively, using *D3* feature vector. The very high % EER of MFCC is due to the fact that it could not identify *S10* spoof (*~40* % EER) as compared to the rest of the features (*11-15* % EER). However, MFCC is shown to have best EER for known attacks. Therefore, considering the fusion factors obtained from the development set, the SBAE features are fused at score-level with other system-based features as shown in Table 4.11.

Results of score-level fusion: Considering score-level fusion of SBAE with MFCC and the cochlear-based features, the average results and the performance with respect to the individual attacks is shown in Table 4.11. The individual average EER using *D3* feature vector for SBAE, MFCC, CFCC, CFCCIF, CFCCIFS on the evaluation data is *2.48* %, *4.25* %, *1.765* %, *2.095* % and *1.6* %, respectively. On fusing SBAE and MFCC features, the average % EER reduces to *1.93* %. On fusion, the performance of known attacks improves due to MFCC and that of the unknown due to SBAE features. Likewise, on fusing with cochlear-based features at $\alpha_f = 0.4$ the average EER was around *1.2* to *1.5* %. As compared to fusion with MFCC, the score-level fusion of SBAE with cochlear-based features improved the performance of unknown attacks, especially for *S10* for which the EER reduced from around *15* % for SBAE and *11.7* % for CFCCIFS (Table 4.6) to *8.97* %. The case of using equal

contribution of both feature set was used as shown in Table 4.11 with $\alpha_f = 0.5$. Such a combination improved the performance of known attacks with only *0.25* % EER. However, at the cost of large % EER for *S10* spoof. For cochlear-based features, their additional contribution was better for known attacks. The performance degraded slightly for *S10* spoof, however, this degradation was minor and did not affect much the average % EER. The overall best average % EER was *1.226* obtained by fusion of SBAE with CFCCIFS which is *50* % improvement over SBAE and *23* % improvement over CFCCIFS features. Thus, on an average, the CFCCIFS perform better over rest of the system-based features.

Table 4.11: EER (in %) in terms of individual attacks, average known attacks, average unknown attacks, average with and without *S10* spoof for SBAE along with their score-level fusion with MFCC, CFCC, CFCCIF and CFCCIFS feature sets using *D3* feature vector at selected α_f on the evaluation set

Feature Sets	Individual Attacks										Average			
	Known					Unknown								
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn	Ukn	w/o S10	Avg.
SBAE: D1	0.41	4.71	0.00	0.00	14.60	10.38	0.39	0.34	3.05	56.96	3.94	14.22	11.85	9.084
SBAE: D2	0.04	3.04	0.00	0.00	3.57	4.07	0.23	0.26	1.01	34.88	1.33	8.09	6.74	4.710
SBAE: D3	0.03	2.99	0.00	0.00	2.26	2.97	0.11	0.52	0.91	15.09	1.06	3.92	3.27	2.488
α_f=0.3														
SBAE+MFCC	0.01	0.93	0.00	0.00	0.82	0.88	0.05	0.02	0.13	16.52	0.35	3.52	2.93	1.935
α_f=0.4														
SBAE+CFCC	0.01	1.21	0.00	0.00	1.49	1.08	0.06	0.03	0.18	9.30	0.54	2.13	1.77	1.335
SBAE+CFCCIF	0.02	1.21	0.00	0.00	1.49	1.27	0.06	0.16	0.34	10.63	0.54	2.49	2.08	1.518
SBAE+CFCCIFS	0.00	0.95	0.00	0.00	1.24	1.00	0.05	0.20	0.27	**8.96**	0.44	2.09	1.75	1.267
α_f=0.5														
SBAE+MFCC	0.00	0.67	0.00	0.00	0.60	0.57	0.04	0.00	0.06	20.87	**0.25**	4.31	3.59	2.28
SBAE+CFCC	0.01	1.07	0.00	0.00	1.49	0.93	0.07	0.01	0.16	9.58	0.52	2.15	1.79	1.333
SBAE+CFCCIF	0.01	0.96	0.00	0.00	1.43	1.13	0.05	0.13	0.23	11.04	0.48	2.52	2.10	1.498
SBAE+CFCCIFS	0.00	0.71	0.00	0.00	1.23	0.82	0.04	0.17	0.21	9.08	0.39	**2.07**	**1.72**	**1.22**

Score-level fusion is carried as per eq. (3.6), Kn=known, Ukn= Unknown, w/o S10=Average without S10, Avg. = Average of S1-S10

Dependency on spoofing algorithms: On the evaluation set, we present the results on the same type, different type and on *S10* spoof separately. For known attacks, the results were similar to that observed in the development set, i.e., these attacks were identified very well. For the same type of attack, as seen in the development set, the both SS and VCS identified speech generated by similar algorithms. For the different type of attacks, again similar to the development set, the VCS spoof identified vocoder-based SS very well. However, the *S3* and *S4* spoofs could not identify VCS spoof that well and the % EER increased with the increase in dynamic information of the features. We consider here the *S10* spoof separately as it is a vocoder-independent spoof and the training is carried on the vocoder-based

speech. The observations for *S10* spoof on training with SS are very much similar to the MFCC and cochlear-based features. That is, the % EER does not decrease with the use of dynamic information and it is as high as *40* %. On the other hand, for VCS, the % EER decreased for *S10* on using the dynamic information. On training with *S1*, the EER is even below *10* %. Thus, appropriate modeling can be done to detect *S10* spoof as well.

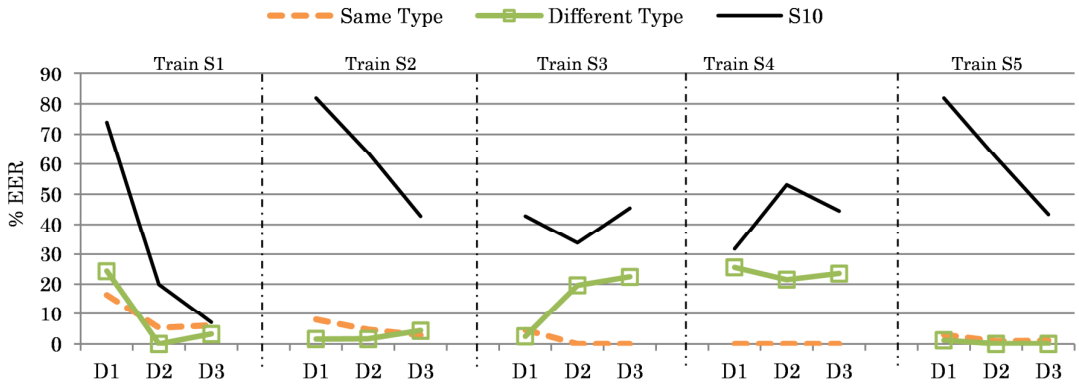


Figure 4.22: The % EER for same type, different type and *S10* attack when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for SBAE feature set using different vectors (i.e., *D1*, *D2*, *D3*) and tested on the evaluation dataset.

Discussion on the DET curves: The DET curves for the MFCC, SBAE and their score-level fusion is shown in Figure 4.23. The DET curve for CFCC and CFCCIFS curves were shown in Figure 4.15. It can be observed from Figure 4.23 (a) that the

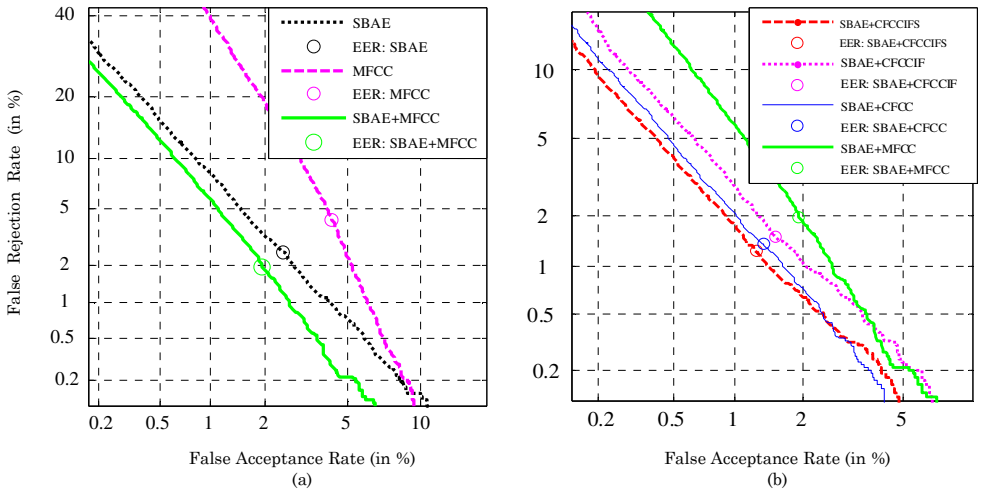


Figure 4.23: The DET curve on the evaluation set for (a) MFCC, SBAE and score-level fusion of MFCC and SBAE feature sets at α_f=*0.3* (b) the score-level fusion on SBAE and MFCC as in (a) and SBAE with cochlear-based features at α_f=0.5.

MFCC had large FRR than SBAE features and slightly better FAR than SBAE at low FRR. However, on fusing both MFCC and SBAE at score-level, the DET curve shows better performance at *all* operating points of the DET curve which indicates that the SBAE features captured complementary information to that of the MFCC features alone. Considering the fusion of SBAE and MFCC features as in Figure 4.23 (a), it is observed in Figure 4.23 (b) that the fusion of SBAE with CFCCIFS gave better performance along all the operating points.

4.6.4.4 Results on the Blizzard Challenge 2012 Database

The results of the performance of the SBAE features on the Blizzard Challenge 2012 dataset are shown in Table 4.12. It was observed on the results of the ASV spoof dataset that the SBAE features showed improvement in the % EER when the dynamic features were used. However, in the case of completely unknown scenarios of channel mismatch, neither static nor the dynamic SBAE features contribute much in detecting the spoofs. Only HMM-based system *E* was observed to be detected with *9* % EER using static features. However, for this % EER increased on using the dynamic information. On the evaluation set of the ASV spoof challenge, the *S10* performance for SBAE was found to be much better as compared to MFCC. However, in the case of Blizzard Challenge 2012 database, the EER for the USS systems was > *40* %. For the case for MFCC as shown in Table 4.7, the results for SBAE are also random without any trend either for USS or HMM-based speech. Just as MFCC, the performance of SBAE features also varies with the use of dynamic features. This shows that on using the system-based features trained on ASV spoof data the detection of statistical and USS-based spoofs may not apply.

4.6.4.5 Results on the Blizzard Challenge 2014 Database

For the Blizzard Challenge 2014 database, the performance of SBAE features is shown in Table 4.12. For the Gujarati language, the HMM-based systems (*D* and *H*) where classified with < *10* % EER while the other HMM-based systems (*C*, *E* and *F*) still had a large % ERR of *10-50* %. For USS-based system *G* the performance improved significantly with the use of dynamic information. In comparison to the SBAE features, the MFCC features show significant decrease in the EER (as shown in Table 4.8). For Hindi, almost all HMM-based systems (except *E*) gave less % EER and the performance of USS-based system *G* degraded with the increase in dynamic

information. Surprisingly, HMM-DNN-based system of Gujarati achieved an EER of *47* % while the Hindi system achieved an EER of *5* %. The consistency of reduction in EER with the dynamic features was not found in the case of SBAE features when evaluated on the ASV spoof challenge database. However, the SBAE features did achieve less % EER for HMM-based speech, which was not the case with MFCC or other cochlear-based features. Thus, these features could not detect extremely unknown attacking scenario and also the features are dependent on the language to a certain extent. However, this could not be justified in the experimental results.

Table 4.12: EER (in %) for SBAE feature set using *D1*, *D2* and *D3* feature vectors on training with the ASV spoof data and testing with Blizzard Challenge databases

Blizzard 2012					Blizzard 2014					Blizzard 2014				
English		SBAE			Gujarati		SBAE			Hindi		SBAE		
		D1	D2	D3			D1	D2	D3			D1	D2	D3
USS	**B**	44	44	40	**C**	**HMM**	61	32	45	**B***	**HMM**	17	2	1
Hybrid	**C**	39	35	38	**D**	**HMM**	9	48	1	**C**	**HMM**	5	0	1
Hybrid	**D***	48	69	73	**E**	**HMM**	59	85	91	**D**	**Hybrid**	26	4	4
HMM	**E***	9	10	23	**F**	**HMM-DNN**	79	47	68	**E**	**HMM**	67	22	30
USS	**F**	70	63	62	**G**	**USS**	84	16	4	**F**	**HMM-DNN**	16	5	6
USS	**G**	23	79	79	**H**	**HMM**	45	8	18	**G**	**USS**	37	43	52
HMM	**H**	14	44	52						**H***	**HMM**	32	21	13
USS	**I**	36	58	40						**K**	**HMM**	3	0	0
Diphone	**J***	56	30	38										
HMM	**K***	43	59	60										

* systems with lower MOS from $1 \leq 2$ (wavefiles for Gujarati system B and system I and Hindi system I are not available)

4.7 Chapter Summary

This Chapter presented the novel combination of envelope and average IF in the CFCC framework and explored a proposed SBAE feature for the task of spoof detection. All the feature sets gave promising results on the ASV spoof 2015 database and especially on the unknown attacks. However, the countermeasures were not robust to the extreme unknown case as that of the systems in the Blizzard challenge datasets. The contribution of the dynamic features significantly depended on the type of spoof and the channel variations as well. These inferences were drawn from system-based features with no component involved explicitly from excitation source-based features. The lack of source information is also a reason for unnaturalness in the spoofed speech. Hence, the next Chapter presents several excitation source features for the SSD task.

Chapter 5.
Source-based Features

5.1 Introduction

In this Chapter, we explore fundamental frequency (F_0) and its various dynamics along with the Strength of Excitation (*SoE*) as various excitation source features for the SSD task. Next, the use of prediction analysis of speech is also carried out. The features derived from Linear Prediction (LP), Long-Term Prediction (LTP) and Non-Linear Prediction (NLP) residual acts as source features for the Spoofed Speech Detection (SSD) task. In addition, we explore the very well-known source-based speech prosody model, i.e., Fujisaki Model to derive features that capture prosodic information in spoofed speech. The source-based features on its own do not contribute to the SSD task. Therefore, these features are fused at score-level with system-based features as discussed in the previous Chapter 4. The combination of source and system features enhances the performance of the SSD system.

5.2 Fundamental Frequency (*Fo*) and Strength of Excitation (*SoE*)

5.2.1 Source Parameter Extraction

The speech signal is not exactly periodic signal [$s(n) \neq s(n+kN)$, where $k \in Z$ and N is the fundamental period], however, short segments of speech are known to exhibit quasi-periodic nature. Hence, rather than having a particular F_0 value, the speech utterance has a time-varying F_0 contour. The F_0 contour of the speech signal actually correlates to the frequency of vibration of the vocal folds. As the vocal folds vibrate, the sudden closure of the folds excites the vocal tract system to generate the speech signal as the output. Thus, the speech production mechanism can be approximated as a Linear Time-Invariant (LTI) system, given by the following representation [**21**],

$$s(t) \approx A\frac{d}{dt}\left[g(t) * h(t)\right] = A\left[\frac{d}{dt}g(t)\right] * h(t), \tag{5.1}$$

$$\therefore \quad s(t) = A\dot{g}(t) * h(t), \tag{5.2}$$

where * is the convolution operation, A is the gain that controls loudness, $s(t)$ is the speech signal, $\dot{g}(t)$ is glottal flow derivative waveform and $h(t)$ is the impulse response of the vocal tract system [**21**]. The derivative effect caused by lip radiation (generally modeled as a first-order differentiator) is considered with source $g(t)$ to give a glottal flow derivative waveform $\dot{g}(t)$ that excites the vocal tract as a system [**177**]. As the vocal folds vibrate, the $g(t)$ shows increase and decrease in the amount of air that passes through the glottis (as shown in Figure 5.1 (a)). There is a cycle-to-cycle variation in the amount of airflow that passes through the folds and duration for which the air passes (when the folds open and close). The duration of one cycle of opening and closing of the vocal fold or the duration from one Glottal Closure Instant (GCI) to another corresponds to a pitch period (i.e., T_0) and the inverse of the time-varying pitch period over various cycles constitutes F_0 contour of the speech signal.

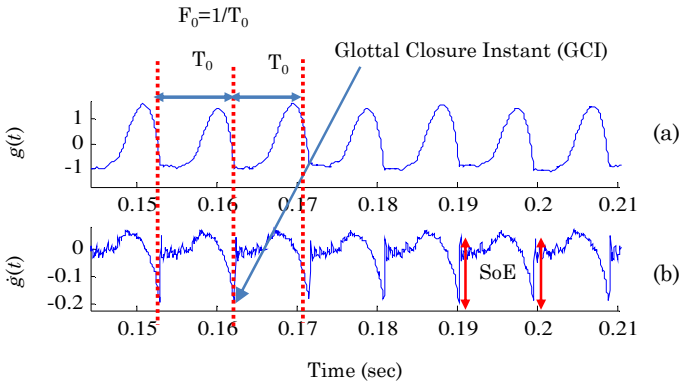


Figure 5.1: (a) The glottal flow waveform ($g(t)$) and (b) glottal flow derivative waveform ($\dot{g}(t)$)

The F_0 contour is known to be the source representation of the speech signal. It carries both *linguistic* and *non-linguistic* information embedded in speech. The voiced speech of a typical adult male will have an F_0 from *85-180* Hz, and that of a typical adult female from *165-255* Hz [**178**]. The vocal folds are heavier for male than for female and hence, they take more time to open and close (in the case of males) which increases the pitch period (T_0) and decreases the fundamental frequency (F_0) [**21**]. In the case of children and infants, the folds are lighter in mass resulting in high pitch (F_0). The change in length of the vocal folds is also responsible for changes in the pitch (F_0). Due to these and several other factors, the rate at which the glottis closes varies. This rate of sudden closure results in impulse-like excitation as observed in the $\dot{g}(t)$. The amplitude of the negative peaks of $\dot{g}(t)$ is known as the *SoE* with which the glottis closes suddenly.

5.2.2 Basis of using F_0 and *SoEs*

Humans vary their vocal fold movements and *SoE* at the glottis depending on the type of utterance and the situation which can affect the F_0 contour and the *SoE* of the speech signal. Therefore, there is some correlation or similarity between the *SoE* estimated at the glottis and that estimated from the speech. In addition, these *SoEs* are also correlated in some manner to the F_0 (as shown later in Figure 5.5). In the case of machine-generated speech, there is *no* true glottal closure phenomenon during generation of the speech signal (especially, for the HTS-based Synthetic Speech (SS) and Voice Converted Speech (VCS)). There are various approaches used to provide excitation source in a vocoder during speech generation. This can be similar to mixed excitation model in which both periodic and aperiodic components are used during the production of speech sounds. While using excitation information, it is necessary to adapt the periodic waveform according to the speaker's F_0 range to sound like the intended speaker. In addition to the F_0, the *SoE* of the periodic waveform, or the envelope of the periodic waveform, i.e., the excitation source will also affect the speech quality in some sense. In the case of natural speech, the *SoE* at the glottis estimated by negative peaks of the $\dot{g}(t)$ and the *SoE* estimated from speech are found to correlate with each other [**179**]. Such an analysis for spoofed speech or synthetically generated speech does not exist. Hence, in this study, we initially assume that for vocoded speech a correlation between the *SoE* at the input to the vocal tract may not exist. To quantify this, we study the effect for vocoded speech and understand the correlation between the F_0 and *SoE* as well. Thus, we have the F_0 contour estimated from the speech signal at the GCIs, and at those GCI locations, the *SoE* is estimated from speech (referred to as *SoE1*) and the *SoE* is estimated from $\dot{g}(t)$ (referred to as *SoE2*).

5.2.3 F_0 and *SoE* Extraction from Speech

To estimate the F_0 contour, the location of the *sudden* closure of the glottis or the GCIs needs to be estimated. Once the GCI locations are estimated, the reciprocal of difference in time between two GCI's constitutes the F_0 contour. In this thesis, we use the Zero Frequency (ZF) filtering method to estimate GCI locations. The main motivation to use ZF filtering approach is that it estimates both F_0 and *SoE* in the same framework. Recently, in [**180**]- [**181**], the authors propose an approach to

estimate the F_0 and *SoE* in the same framework. However, this uses thresholding and may not be accurate enough and hence, we use the ZF-filtering approach. The ZF filter is a digital second order resonator with complex conjugate poles $p_1 = re^{j\omega_0}$ and $p_2 = p_1^* = re^{-j\omega_0}$ near the unit circle. The Z-domain system function of the resonator is given as,

$$H(z) = \frac{b_o}{(1 - p_1 z^{-1})(1 - p_2 z^{-1})}. \tag{5.3}$$

In eq. (5.3), when the poles are near to the unit circle (i.e., $r \to 1$) and makes an angle of ω_o with the positive direction of X-axis, then the resonant frequency ω_r is given as,

$$\omega_r = \cos^{-1}\left[\frac{(1+r^2)}{2r}\cos\omega_0\right], \tag{5.4}$$

In eq. (5.4) when $r \to 1$, the resonant frequency ω_r equals to the pole angle ω_o on the unit circle. As the ZF filter resonates at *0*-Hz, the pole angle $\omega_0 \approx \omega_r \approx 0$ and hence, $p_1 = p_2^* = r$, and eq. (5.3) can be written as,

$$H(z) = \frac{1}{(1 - z^{-1})(1 - z^{-1})} = \frac{1}{(1 - z^{-1})^2}. \tag{5.5}$$

Using the above response to filter the speech signal removes the high-frequency components and leaves the low-frequency components along with the d.c. bias. To remove the bias, trend removal is carried out on the ZF filtered signal $y[n]$ over an analysis window to get a sinusoidal-like signal $y[n]$, i.e.,

$$y[n] = y_1[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_1[n+m], \tag{5.6}$$

where $2N+1$ is the samples corresponding to the trend removal window. Thus, the ZF filtered signal consists of both ZF filtering and trend removal process. The ZF filtering is based on the concept that the effect of an impulse is spread uniformly across *all* the frequencies including the zero frequency. For speech signals, the excitation is impulse-like and hence, the output of ZF filter will give an estimate of the epoch locations (i.e., GCIs). A detailed description of the ZF filtering approach is given in [**182**]. Therefore, to estimate the GCI locations, the ZF filtering is performed on the speech signal and the negative-to-positive zero-crossings of the filtered signal are hypothesized as an estimate of GCIs [**182**]. The slope of ZF filtered signal at

negative-to-positive zero-crossings gives a measure of the strength of glottal closure, i.e., *SoE* [**179**]. Figure 5.2 shows an illustration of the F_0 and *SoE* estimated from the speech signal using the ZF filtered signal. It is observed that for the given speech signal, the F_0 and *SoE* follow a nearly similar pattern. That is, there is some correlation between the F_0 and *SoE* for the natural speech signal. Therefore, we explore such correlations between spoofed speeches for the SSD task.

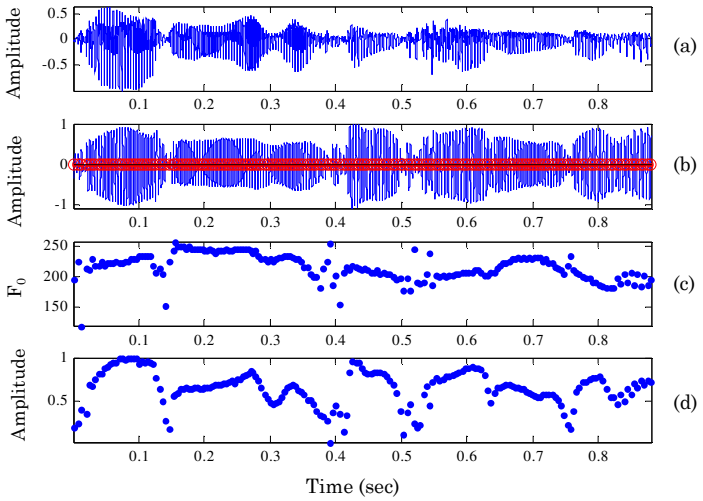


Figure 5.2: (a) Voiced regions of a speech signal (b) ZF filtered signal (c) F_0 contour from GCI locations (negative-to-positive zero-crossings of (b)) and (d) *SoE* at GCIs (slope at negative-to-positive zero-crossings of (b)).

5.2.4 Estimation of Glottal Flow Waveform ($g(t)$)

To obtain an initial estimate of the $g(t)$, we use the Iterative Adaptive Inverse Filtering (IAIF) method to decompose speech into its glottal source signal and vocal tract system [**183**]. The block diagram of IAIF method is shown in Figure 5.3 [**184**].

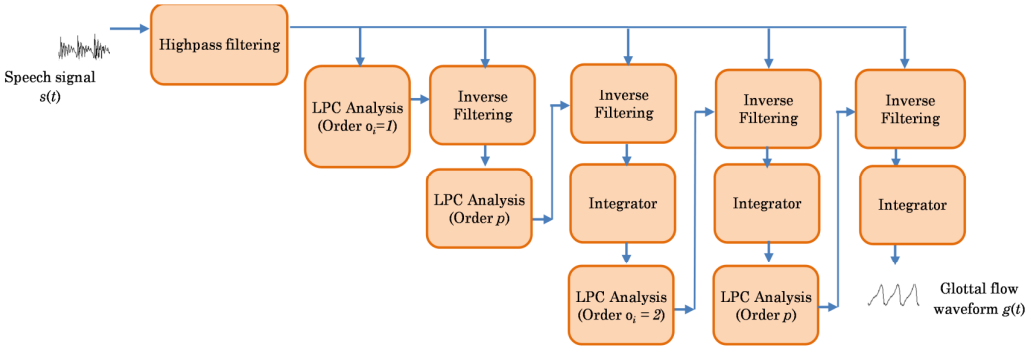


Figure 5.3: Block diagram of the IAIF method. Adapted from [184**].**

In the IAIF method, the effect of the vocal tract system and lip radiation is canceled from the speech signal to give an estimate of $g(t)$. The key motivation for using the IAIF method is that there is no need of the ground truth (such as electroglottograph (EGG)) and it is computationally efficient and completely automatic. To obtain an initial estimate of the glottal flow, the values of Linear Prediction Coefficients (LPC) are taken as $o_i=1$ and $o_i=2$ for the first and second iteration, respectively. In both iterations, to obtain the resonances of the vocal tract, the pole order $p=20$ is considered for an $Fs=16$ kHz (due to the relationship between sampling frequency Fs and length of vocal tract [**185**]). From estimated $g(t)$, its derivative is considered to obtain $\dot{g}(t)$, which can be further used to estimate the *SoE* from $\dot{g}(t)$.

Estimation of *SoE* from $g(t)$: Once the $g(t)$ is estimated from the speech signal, its derivative is computed and the amplitude of the negative peak of the $\dot{g}(t)$ at each GCIs is hypothesized as the *SoE* as shown in Figure 5.1 (a). The following sub-Sections discuss the detailed analysis and extended results to that presented in [**86**].

5.2.5 Analysis of F_0, *SoE1* and *SoE2* on Spoofed Speech

In this Section, we observe the differences between natural and spoofed speech for F_0, *SoE1*, and *SoE2*. Figure 5.4 shows the F_0 and *SoE1* derived from speech and

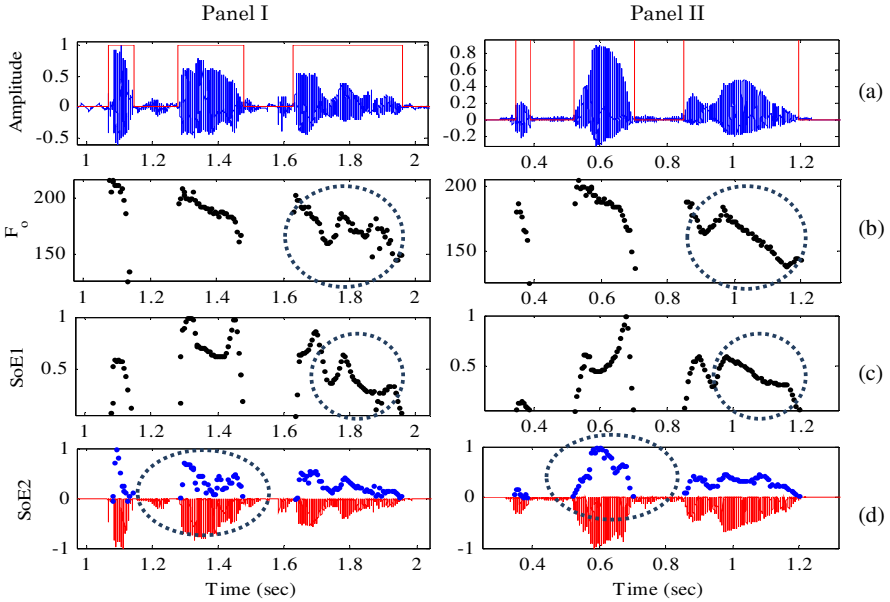


Figure 5.4: Panel I: Natural speech and Panel II: vocoder-based SS: (a) speech signal \It's nice to hear\, (b) F_0 contour estimated by ZF filtering (c) normalized *SoE1* at GCIs estimated by ZF filtering and (d) the $\dot{g}(t)$ (red) and normalized *SoE2* estimated from $\dot{g}(t)$ at GCIs estimated from ZF filtering (dotted blue). Adapted from [86].

SoE2 derived from the $\dot{g}(t)$ for a natural speech (Panel I) and HMM-based SS spoof (Panel II) from the SAS database [**19**]. In Figure 5.4 (d), only the negative part of the $\dot{g}(t)$ is plotted and the magnitude of $\dot{g}(t)$ at the GCI is indicated as *SoE2* in Figure 5.4 (d). As shown by the dotted regions in Figure 5.4, there exist variations in excitation source features for natural and SS speech. The F_0 contour of natural speech had more variations as compared to that of the SS speech (i.e., more dynamic information of the F_0 contour of natural speech as compared to the SS speech in Figure 5.4). These variations were even observed in the *SoE* estimated from speech and the $\dot{g}(t)$. For this particular case of spoofed speech, the variations were less in F_0, *SoE1* and *SoE2* as compared to natural speech. Similar variations were observed over several utterances for vocoder-based SS and VCS spoof.

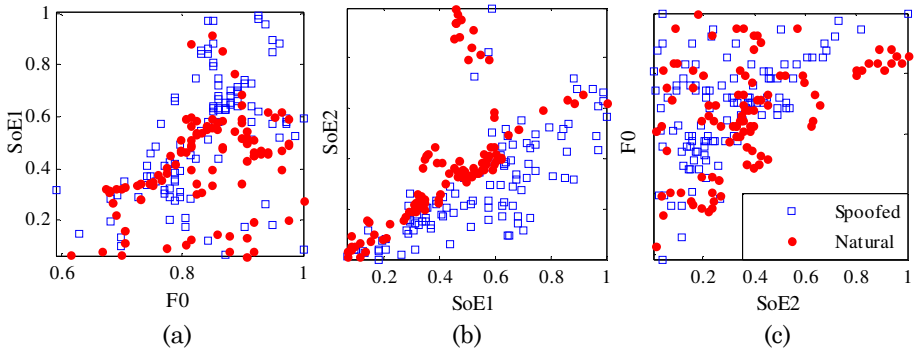


Figure 5.5: Scatterplots for (a) *F_0 vs. SoE1* (b) *SoE1 vs. SoE2* and (c) *SoE2 vs. F_0* for the natural and vocoder-based SS utterance in Panel I and Panel II, respectively (from Figure 5.4). Adapted from [86].

The relation between source-based features for natural and SS spoof in Figure 5.4 is shown by scatter plot of F_0, *SoE1,* and *SoE2* at GCIs in Figure 5.5. The correlation coefficients (for the speeches shown in Figure 5.4) between F_0 *vs. SoE1, SoE1 vs. SoE2* and *SoE2 vs.* F_0 are *0.51*, *0.73* and *0.51* for natural speech and *0.34*, *0.645* and *0.45* for SS speech, respectively. Thus, it is observed that *correlations* vary for natural and SS speech. Although a direct relationship amongst F_0, *SoE1* and *SoE2* cannot be specified for different spoofing algorithms, there do exist differences in natural and spoofed speech due to the excitation source characteristics. This will be verified by using F_0, *SoE1* and *SoE2* and their dynamics as discriminative features for SSD task.

5.2.6 Experimental Results

5.2.6.1 Parameterization

The source features, i.e., F_0, *SoE1,* and *SoE2* are extracted at GCIs estimated by ZF and IAIF method, respectively, using a frame size of *25 ms* and with a frame shift of *50 %* (after discarded the unvoiced regions). The F_0, *SoE1,* and *SoE2* give a *3*-dimension *(3-D)* static feature vector, i.e., *Ds* for each GCI location. The dynamics of the F_0, *SoE1,* and *SoE2* features are also considered by taking their first derivative, i.e., velocity, (*d1*: ΔF_0, Δ*SoE1,* and Δ*SoE2*) and appended to the *Ds* to get *6-D* feature vector (*D1=Ds+d1*). This was done till 5^{th} order derivative (i.e., acceleration, jerk, jounce, crackle) to get *D2*, *D3*, *D4,* and *D5*, corresponding to *9-D*, *12-D*, *15-D* and *18-D* feature vectors, respectively. For using system-based information, the score-level fusion of *36-D* system-based MFCC, CFCC, CFCCIFS and SBAE feature vectors comprising of static and dynamic information (i.e., *12*-static+*12*-Δ +*12*-$\Delta\Delta$) are used.

5.2.6.2 Results on the Development Set of ASVspoof challenge Database

Effect of source features and their dynamics: The source-based features and their dynamics are lesser in dimensions. Therefore, the effect of these source features is studied by evaluating the % EER of the detector for various number of mixture components in GMM (as shown in Figure 5.6). The testing is done on the development set for the models trained on the training data.

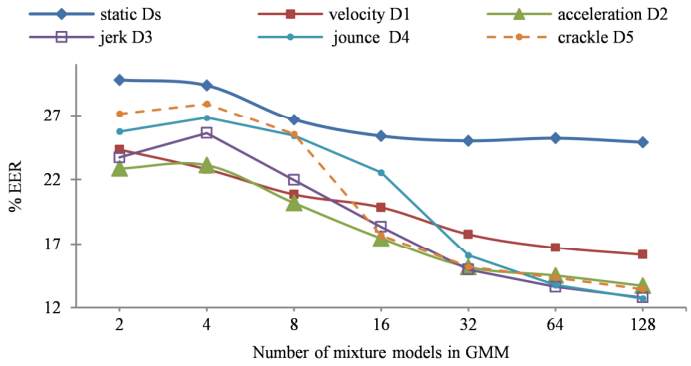


Figure 5.6: The % EER obtained on the development set when the static and various dynamics, i.e., velocity, acceleration, jerk, jounce and crackle of F_0, *SoE1* and *SoE2* are considered. Adapted from [86].

It is observed from Figure 5.6 that the % EER on the development set decreases significantly when the dynamic information is added to the static features. The

decrease in % EER is significant with the increase in the number of mixture models in GMM. The EER for *Ds*, *D1*, *D2*, *D3*, *D4* and *D5* are *24.8* %, *16.1* %, *13.6* %, *12.7* %, *12.6* %, and *13.4* %, respectively, on using *128* mixtures. With higher-order derivative than the jerk (*D3*), the decrease is not significant and also increases slightly. Thus, the *D3* feature vector with *128* mixture GMM can be considered.

To observe the effect of F_0, *SoE1* and *SoE2*, the % EER with F_0, *SoE1* and *SoE2* used individually up to third order derivative (i.e., *12-D*) was estimated. Next, the performance of using only two features at a time is also studied. It is observed from Table 5.1 that individually for F_0, *SoE1* and *SoE2* features, the % EER is very high *~27 %*. On a feature-level fusion of the *D3* feature vector of F_0 features with two *SoEs* one at a time, the % EER increased. However, on combining the two *SoEs* with the F_0, the % EER decreased significantly (indicating that the *SoEs* capture complementary information). As indicated in Section 5.2.2, while synthesizing or converting speech, the information of F_0 is provided whereas the *SoE* is not explicitly provided and hence, the use of *SoEs* gave better performance for the SSD task. However, the EER is not less than *12.7 %* that was obtained when all the three features are used (as shown in Figure 5.6). Thus, all F_0, *SoE1,* and *SoE2* features are essential for detecting spoofed speech.

Table 5.1: EER (in %) for F_0, *SoE1* and *SoE2* feature set used alone and when combined with each other using *D3* feature set. Adapted from [86]

Individual Feature Set	% EER	Feature-level Fusion	% EER
D3: F_0	27.94	D3: F_0 & SoE1	45.98
D3: SoE1	25.54	D3: F_0 & SoE2	43.92
D3: SoE2	27.68	D3: SoE1 & SoE2	**18.82**

Fusion with the system-based features: The score-level fusion of source-based features with the various system-based features is shown in Table 5.2. It was observed that the % EER of source-based features when used alone, did not decrease much after *D3* feature vector. However, on fusing with the system-based features, the % EER decreased for *D4* and *D5* feature vector as well. It was observed that for $\alpha_f = 0.8$ for all system-based features, the % EER of the after score-level fusion is minimum. The MFCC, CFCC, CFCCIF, CFCCIFS and SBAE features achieved the best % EER of *0.94, 0.80*, *0.66*, *0.54* and *1.12* with *D5*, *D5, D4, D5 and D2* features vectors of F_0, *SoE1* and *SoE2*. It is observed that on score-level fusion, the % EER of system-based features reduced to almost half (except SBAE). The least EER among

these combinations is *0.54 %* which is less than that submitted for the ASVspoof 2015 challenge (i.e., *0.83 %* with the fusion of MFCC and CFCCIF). Thus, considering the best feature vector as *D5* with a fusion factor of $a_f = 0.8$ for source-based features, the % EER is obtained on the evaluations set.

Table 5.2: EER (in %) for *F0*, *SoE1,* and *SoE2* features using all feature vectors and their score-level fusion with system-based feature sets (using *D3* feature vector) at various fusion factors a_f on the development set

Feature Set 1	Fusion Factor (a_f)											Feature Set 2
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
Ds	24.88	9.87	5.46	3.43	2.14	1.66	1.40	1.40	1.46	1.52	1.60	
D1	16.13	8.32	4.92	3.15	1.94	1.43	1.17	1.20	1.32	1.43	1.60	
D2	13.70	7.84	4.86	3.03	1.83	1.26	1.06	1.09	1.17	1.34	1.60	**MFCC**
D3	12.78	7.55	4.98	3.17	1.92	1.26	1.06	**1.03**	1.09	1.29	1.60	**12s+Δ+ΔΔ**
D4	12.70	7.95	5.49	3.57	2.26	1.49	1.09	0.97	**1.03**	1.29	1.60	
D5	13.44	8.95	6.35	4.18	2.83	1.77	1.20	1.03	**0.94**	1.20	1.60	
Ds	24.88	9.24	4.66	2.77	1.97	1.52	1.26	1.29	1.37	1.43	1.54	
D1	16.13	7.95	4.43	2.55	1.77	1.34	1.03	1.00	1.12	1.34	1.54	
D2	13.70	7.49	4.26	2.77	1.69	1.14	0.89	0.86	0.97	1.23	1.54	**CFCC**
D3	12.78	7.38	4.43	2.83	1.74	1.29	0.92	0.86	**0.83**	1.14	1.54	**12s+Δ+ΔΔ**
D4	12.70	7.75	4.95	3.03	1.92	1.34	1.00	0.86	**0.80**	1.06	1.54	
D5	13.44	8.86	5.95	3.66	2.37	1.54	1.14	0.86	**0.80**	1.03	1.54	
Ds	24.88	12.01	6.03	3.52	2.40	1.69	1.20	1.17	1.20	1.34	1.52	
D1	16.13	9.55	5.58	3.26	2.20	1.46	1.03	0.89	0.97	1.23	1.52	
D2	13.70	8.78	5.29	3.29	2.14	1.29	0.89	0.71	0.80	1.12	1.52	**CFCCIF**
D3	12.78	8.58	5.55	3.46	2.17	1.52	1.06	0.77	**0.71**	1.00	1.52	**12s+Δ+ΔΔ**
D4	12.70	8.98	5.98	3.92	2.26	1.77	1.20	0.74	**0.66**	0.94	1.52	
D5	13.44	10.09	7.21	4.80	3.15	2.00	1.26	0.89	**0.69**	0.89	1.52	
Ds	24.88	11.27	5.43	3.03	2.00	1.37	1.03	0.92	0.92	1.06	1.23	
D1	16.13	9.21	5.03	2.89	1.94	1.12	0.80	0.69	0.83	0.94	1.23	
D2	13.70	8.46	4.95	3.00	1.77	1.06	0.66	0.57	0.74	0.86	1.23	**CFCCIFS**
D3	12.78	8.32	5.03	3.12	1.94	1.14	0.77	0.66	**0.60**	0.83	1.23	**12s+Δ+ΔΔ**
D4	12.70	8.75	5.55	3.52	2.03	1.37	0.83	0.66	**0.57**	0.80	1.23	
D5	13.44	9.87	6.78	4.26	2.69	1.60	1.00	0.71	**0.54**	0.74	1.23	
Ds	24.88	11.07	6.69	4.23	3.03	2.26	1.69	1.43	1.34	1.40	1.49	
D1	16.13	9.38	6.15	4.15	2.92	2.06	1.52	1.29	1.23	1.32	1.49	
D2	13.70	8.58	5.86	3.95	2.77	1.94	1.46	1.23	**1.12**	1.23	1.49	**SBAE**
D3	12.78	8.41	6.03	4.20	2.97	2.06	1.52	1.34	**1.14**	1.23	1.49	**12s+Δ+ΔΔ**
D4	12.70	8.81	6.41	4.52	3.06	2.23	1.80	1.40	**1.17**	1.20	1.49	
D5	13.44	9.67	7.43	5.43	3.77	2.75	1.97	1.49	1.26	**1.20**	1.49	

Score-level fusion is carried as per eq. (3.6)

Dependency on spoofing algorithms: To check the discriminative property of the proposed feature set in terms of the dependency to the spoofing algorithm, the systems were trained on individual spoofs and tested on all the spoofs of the development set. As discussed earlier, we further split unknown attacks into two categories, namely, 'same type' and 'different type'. For example, for *S1* VCS spoof: testing with *S1* itself is 'known', testing with its similar kind (i.e., VCS, *S2* and *S5*) is the 'same type' and testing with a different class (i.e., SS, *S3,* and *S4*) is 'different type'. Average of the same type and different type constitutes 'unknown' attacks.

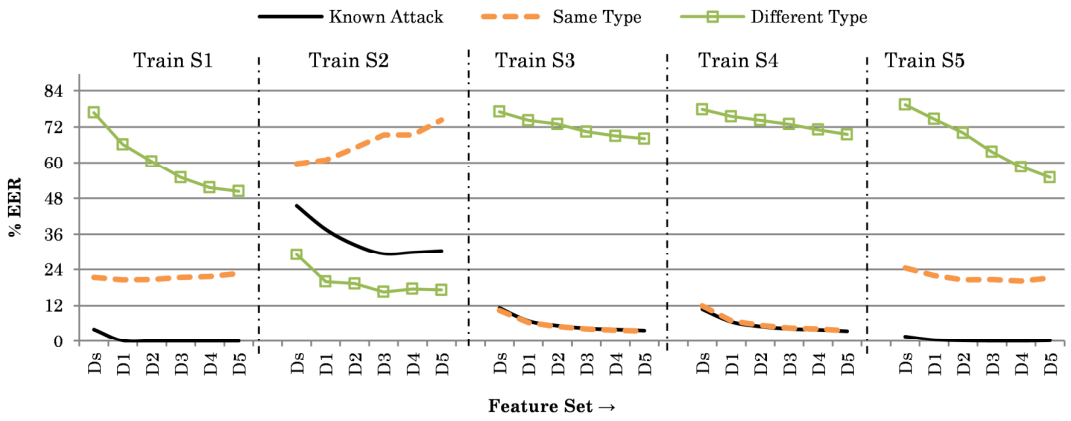


Figure 5.7: The % EER for known, same and different type of attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for *F0*, *SoE1,* and *SoE2* feature set using its various dynamics and tested on the development dataset.

As shown in Figure 5.7, when tested on individual spoofs, the trend is similar for all spoofs except *S2*. Excluding *S2* from the discussion, for the known type of attacks, the % EER decreases with the increase in the dynamic information of *F0*, *SoE1,* and *SoE2* features. That is, the *D5* feature vector works well to classify the natural *vs.* spoofed speech for models are trained on one spoof and tested on the same spoof. Next, considering the 'same type' of attacks in testing, it is observed that with the increase in dynamic information of *F0*, *SoE1,* and *SoE2* features even for the same type of attacks, the % EER is very high for the VCS case. In other words, *S2* and *S5* spoof when tested on system trained by *S1*, then a high % EER was noted. Unlike the case of known attacks, the % EER did not decrease gradually with the use of dynamic information. For the SS spoof, the *S3* and *S4* attacks are developed with the same algorithm and the difference which exists between them is the amount of data used for training. Therefore, for the SS spoofs on testing with its same type, the % EER is almost similar to the known case. Thereafter, for the 'different type' of attacks, the % EER is high for all *S1-S5* attacks. However, the % EER decreases with the increase in dynamic information in the speech signal. The *S2* spoofing attack shows a different behavior for known and same type of attacks as compared to other spoofing algorithms. On testing with itself, the *S2* spoof could not detect itself, i.e., a very high *30* % EER was obtained even with the *D5* feature vector. For the same type of attacks (i.e., VCS), the % EER was found to increase with the increase in the dynamic information. On the other hand, the *S2* spoof could identify SS spoofs better than the other VCS spoofs (*S1* and *S5*). The results for source-based features

on the known and unknown attack are not as promising as compared to the system-based features. However, complementary spoof-specific information in the features can be observed due to dynamic variability between natural and spoofed speech.

5.2.6.3 Results on the Evaluation Set of ASVspoof challenge Database

The results of the evaluation set for source-based features and score-level fusion with system-based features are shown in Table 5.3. It is known from results of the development set that *Ds-D2* feature vectors does not contribute much to the decrease in EER. The % EER decreases from *Ds* to *D4* and increases slightly for the *D5* feature vector. In this case, the performance of *S10* spoof was decreasing. However, the % EER of other vocoder-based spoof increased, giving high average % EER.

Table 5.3: EER (in %) in terms of individual attacks, average known attacks, average unknown attacks, average with and without *S10* spoof for the F_0, *SoE1,* and *SoE2* feature set using *Ds* to *D5* feature vectors and score-level fusion of *D3-D5* feature vectors with the system-based feature set (using *D3* feature vector) at selected α_f on the evaluation set

Feature Sets	Individual Attacks										Average			
	Known Attacks					Unknown Attacks								
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn	Ukn	w/o S10	Avg
Ds	16.40	57.46	25.30	24.10	10.10	19.80	16.50	13.84	24.13	56.63	26.67	26.18	23.07	26.43
D1	2.66	55.76	11.40	10.80	2.82	8.59	7.11	5.86	10.25	61.29	16.69	18.62	12.81	17.65
D2	0.07	54.96	8.13	7.95	0.90	3.40	2.68	1.82	4.23	56.55	14.40	13.74	9.35	14.07
D3	0.01	53.90	6.35	6.34	0.23	1.58	0.86	0.58	2.99	51.25	13.37	11.45	8.09	12.41
D4	0.01	54.84	7.54	8.07	0.20	1.91	0.63	0.71	3.70	40.96	14.13	9.58	8.62	11.85
D5	0.01	54.59	9.34	10.05	0.14	1.39	0.76	0.78	4.03	39.81	14.82	9.35	9.01	12.09
MFCC	0.01	1.04	0.00	0.00	0.86	0.94	0.05	0.00	0.09	37.80	0.38	7.78	0.31	4.08
CFCC	0.00	1.40	0.00	0.00	2.30	1.00	0.10	0.10	0.20	12.30	0.74	2.74	0.57	1.74
CFCCIF	0.03	0.72	0.00	0.00	2.24	0.98	0.16	0.88	0.29	15.42	0.60	3.55	0.59	2.07
CFCCIFS	0.00	0.50	0.00	0.00	1.70	0.70	0.10	1.00	0.20	11.70	0.44	2.74	0.47	1.59
SBAE	0.03	2.99	0.00	0.00	2.26	2.97	0.11	0.52	0.91	15.09	1.06	3.92	1.09	2.49
D3+MFCC	0.00	0.72	0.00	0.00	0.19	0.30	0.02	0.00	0.03	34.47	0.18	6.96	0.14	3.57
D4+MFCC	0.00	0.68	0.00	0.00	0.11	0.24	0.02	0.00	0.02	33.03	0.16	6.66	0.12	3.41
D5+MFCC	0.00	0.69	0.00	0.00	0.08	0.20	0.02	0.00	0.02	32.32	0.15	6.51	**0.11**	3.33
D3+CFCC	0.00	1.13	0.00	0.00	0.68	0.45	0.08	0.02	0.12	11.51	0.36	2.43	0.28	1.40
D4+CFCC	0.00	1.23	0.00	0.00	0.56	0.41	0.07	0.02	0.10	11.74	0.36	2.47	0.27	1.41
D5+CFCC	0.00	1.20	0.00	0.00	0.45	0.35	0.07	0.02	0.08	11.74	0.33	2.45	0.24	1.39
D3+CFCCIF	0.00	0.73	0.00	0.00	0.40	0.31	0.05	0.36	0.08	15.15	0.23	3.19	0.21	1.71
D4+CFCCIF	0.00	0.68	0.00	0.00	0.28	0.27	0.03	0.30	0.06	15.02	0.19	3.14	0.18	1.66
D5+CFCCIF	0.00	0.99	0.00	0.00	0.20	0.23	0.03	0.25	0.07	15.48	0.24	3.21	0.20	1.73
D3+CFCCIFS	0.00	0.43	0.00	0.00	0.29	0.18	0.04	0.42	0.05	11.29	0.14	2.40	0.16	1.27
D4+CFCCIFS	0.00	0.46	0.00	0.00	0.18	0.16	0.03	0.33	0.05	11.37	**0.13**	**2.39**	0.13	**1.26**
D5+CFCCIFS	0.00	0.55	0.00	0.00	0.13	0.11	0.03	0.24	0.07	11.70	0.13	2.43	0.13	1.28
D3+SBAE	0.01	2.76	0.00	0.00	0.52	1.23	0.07	0.22	0.44	14.41	0.66	3.27	0.58	1.97
D4+SBAE	0.01	2.88	0.00	0.00	0.38	1.15	0.05	0.18	0.43	14.59	0.65	3.28	0.56	1.97
D5+SBAE	0.00	3.08	0.00	0.00	0.26	0.97	0.05	0.14	0.41	14.52	0.67	3.22	0.55	1.94

Score-level fusion is carried as per eq. (3.6), Kn=known, Ukn= Unknown, w/o S10=Average without S10, Avg. = Average of S1-S10

We consider score-level fusion for *D3-D5* with system-based features at $\alpha_f = 0.8$ (i.e., the weight of fusion α_f is optimized w.r.t the performance of the SSD system). For fusion with MFCC, the average % EER decreases from *D3* to *D5* feature set. For CFCC and SBAE features, the % EER is almost constant using any of *D3* to *D5*

feature vector. For CFCCIF and CFCCIFS, the same pattern is observed and both gave better EER of *1.66* % and *1.26* % with *D4* feature vector, respectively. Thus, F_0, *SoE1* and *SoE2* features contribute to decrease in % EER than using the system-based features alone.

Dependency on spoofing algorithms: Just as in the case of development set, the testing of the entire evaluation data is carried out when trained on individual *S1*, *S2*, *S3*, *S4* and *S5* spoofs. For the development set, known and same types of spoofs were easily identified. As shown in Figure 5.8, when trained on individual spoofs, the trend is similar for all the spoofs just as the case in development set. The known case is not shown here because it was similar to that of the development set, i.e., for the known type of attacks, the % EER decreases with the increase in the dynamic information of F_0, *SoE1,* and *SoE2* features. The *S3* and *S4* detected itself with ~ *2* % EER while the *S1* and *S5* detected itself with ~ *0* % EER and *S5* with ~ *30* % EER. Next, considering the 'same type' of attacks, *S3* and *S4* detected each other with an EER of ~ *2* %. On the other hand, the *S1* spoof detected *S2* with *50* % EER, *S5* with *0.5* % EER and *S6* to *S9* within *4-8* % EER. Likewise, *S5* detected *S1* with almost *0* % EER, *S6-S9* with *1-5* % EER and *S2* with ~*50* % EER. Therefore, the *S2* spoof tends to increase the average % EER for the same type of attack when trained on *S1* and *S5*. Thereafter, for the 'different type' of attacks, the EER (*50* % - *70* %) is high for all *S1-S5* attacks (except *S2*). However, the % EER decreases with the increase in dynamic information in speech. On the other hand, the *S2* spoof could identify SS spoofs better (~*18* % EER) than the other VCS spoofs (*S1* and *S5*) with > *60* % EER and other *S6-S9* spoofs with *15-25* % EER. Hence, the *S2* spoof does not detect its same type as good as *S1* and *S5*. The model trained on *S2* spoof detected *S3-S4* much better and hence, the average of different attack is less for *S2*. Now, considering the testing results of *S10* vocoder-independent spoof separately, it is observed that the *S10* spoof was detected with *S1* and *S5* with ~ *35* % EER. In addition, the % EER decreased with the increase in dynamic information. On the other hand, for training with the *S2*, *S3* and *S4* spoof, the EER increases with the dynamic information and reaches to about *60* %. The *S2* spoof was generated by transforming the first coefficient of the source speaker's MCC (c1) by a linear transformation to that of the target speaker. In [**123**], the *S2* spoof had less spoof detection error rate and was

more close to natural. However, its effect on the F_0, *SoE1,* and *SoE2* features that led to this different behavior could not be identified.

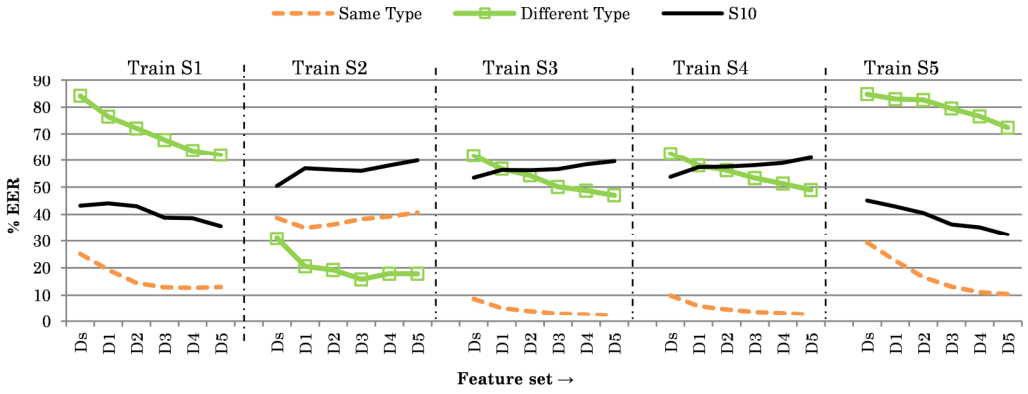


Figure 5.8: The % EER for the same type, different type and *S10* attack when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for *F₀*, *SoE1,* and *SoE2* feature set using its various dynamics and tested on the evaluation dataset.

Discussion on the DET curves: The DET curves for source-based features using *D5* feature vector, MFCC, CFCC, CFCCIF and SBAE features is shown in Figure 5.9 (a). It is seen that source-based features (*D5*) had high % FRR and % FAR. On the other hand, CFCC, CFCCIFS and SBAE features had significantly low % FRR than MFCC. On score-level fusion with the source-based features, the % FRR reduces further which contributes to the decrease in % EER (as shown in Figure 5.9 (b)). The fusion of *D5* source-based features with CFCCIFS gives relatively best performance than the other system-based features.

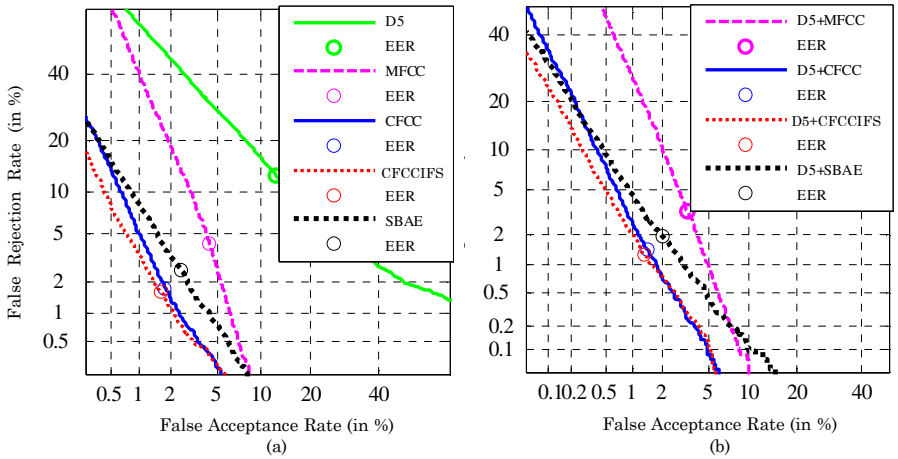


Figure 5.9: DET curves on the evaluation set for (a) *D5* source-based feature (green), MFCC (magenta), CFCC (blue), CFCCIFS (red) and SBAE (black)), (b) score-level fusion *D5* with MFCC (magenta), *D5* with CFCC (blue), *D5* with CFCCIFS (red) and *D5* with SBAE (black) all at $\alpha_f = 0.8$.

5.2.6.4 Results on the Blizzard Challenge 2012 Database

The results of the source-based features from *Ds* to *D5* are shown in Table 5.4 where it is observed that the % EER reduces from *Ds* to *D5* for all the systems from *B* to *K*.

Table 5.4: EER (in %) for *F_0*, *SoE1* and *SoE2* feature set using *Ds* to *D5* feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2012 database

Blizzard 2012	Systems	Feature Sets					
		Ds	D1	D2	D3	D4	D5
USS	**B**	39	34	36	32	33	**30**
Hybrid	**C**	55	61	57	54	55	**49**
Hybrid	**D***	61	56	33	16	13	**6**
HMM	**E***	9	10	5	3	3	**1**
USS	**F**	61	59	50	44	40	**38**
USS	**G**	33	24	17	11	10	**8**
HMM	**H**	47	37	26	19	17	**12**
USS	**I**	48	45	38	30	30	**24**
Diphone	**J***	47	40	30	17	16	**10**
HMM	**K***	8	7	3	1	0	**0**

*systems with lower MOS from $1 \leq 2$

The decrease in % EER is more significant for systems with lower MOS. Both system *C* and *D* are hybrid systems, yet the EER are significantly different. For USS-based systems *B*, *F*, *G* and *I*, the % EER is *30*, *38*, *8* and *24*, respectively, using the *D5* feature vector. For the HMM-based systems *E*, *H* and *K*, the % EER are significantly less than either USS-based speech or hybrid systems. The relative improvement in USS-based systems *B*, *F*, *G* and *I* from *Ds* to *D5* feature vector is *23.07* %, *37.70* %, *65.21* % and *50.00* %. These improvements are less as compared to the statistical-based methods *E*, *H* and *K* with a relative improvement of *88.88* %, *74.46* % and *100* %, respectively. Even the diphone-based *J* system showed a relative improvement by *78.72* % decrease in EER. Thus, for the Blizzard Challenge 2012 database, the prosody-based derived from dynamic variation of *F_0*, *SoE1,* and *SoE2* features gave best % EER with *D5* feature vector. Therefore, the source-based features can aid in detecting vocoder-based spoofs, indicating that the *F_0* and *SoEs* features carry crucial differences between natural and spoof speeches.

5.2.6.5 Results on the Blizzard Challenge 2014 Database

The results on the Gujarati language are similar to that of the Blizzard Challenge 2012 database with the % EER decreasing from *Ds* to *D5*. The % EER for all the systems showed significant improvement in the performance. For HMM-based

system *D*, the decrease was only from *32* % to *15* % using *D3* feature vector. For system *F*, the % EER is as low as *0* % for the HMM-DNN-based system.

Table 5.5: EER (in %) for F_0, *SoE1* and *SoE2* feature set using *Ds* to *D5* feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2014 database for the Gujarati language

Blizzard 2014	Gujarati Systems	Feature Sets					
		Ds	D1	D2	D3	D4	D5
HMM	C	57	44	26	9	10	**6**
HMM	D	32	25	19	**15**	19	24
HMM	E	70	55	30	**12**	**12**	**12**
HMM-DNN	F	49	34	6	**0**	1	**0**
USS	G	65	60	46	19	**13**	9
HMM	H	29	30	13	4	**3**	**3**

* wavefiles for baseline system B and system I are not available

Table 5.6: EER (in %) for F_0, *SoE1* and *SoE2* feature set using *Ds* to *D5* feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2014 database for the Hindi language

Blizzard 2014	Hindi Systems	Feature Sets					
		Ds	D1	D2	D3	D4	D5
HMM	B*	56	28	20	14	**12**	16
HMM	C	22	10	**5**	6	**5**	**5**
Hybrid	D	47	40	35	32	**28**	**28**
HMM	E	46	15	6	**4**	**4**	**4**
HMM-DNN	F	33	14	4	**3**	**3**	**3**
USS	G	26	12	7	**4**	**4**	**4**
HMM	H*	2	0	1	**0**	**0**	**0**
HMM	K	9	6	**4**	6	5	7

* systems with lower MOS from $1 \leq 2$ (wavefiles for system I are not available)

Similar observations are found for the % EER estimated on the Hindi language. Except for hybrid approach-based system *D*, the % EER of all the systems decreased significantly from *Ds* to *D5* feature vector. The HMM-based system *B* has low MOS, however, with *Ds* feature vector an EER of *56* % is obtained. On using the dynamic information, the EER reduced to as low as *12* %. It is to be noted that, on an average, the results obtained by the source-based F_0, *SoE1,* and *SoE2* features are better than that obtained by MFCC, CFCC, CFCCIF, CFCCIFS feature sets. Thus, the source-based features, especially, the dynamic information is highly essential in capturing the synthetic nature of the spoofed speech and hence, improving the performance of the SSD system. Moreover, it should be observed that for the Blizzard Challenge 2012 database, the prosody-based features gave best % EER with *D5* feature vector. However, for the Blizzard Challenge 2014 database, the best % EER was not always for the *D5* feature vector. This may be due to the fact that the prosodic features are language-dependent and the dynamic information in source-based features varies

with the language. The dynamic information is essential for the SSD task, however, the amount and nature of dynamics can be dependent on various parameters.

5.2.7 Summary

The main aim of this Section was to showcase that there are dynamic variations in the parameters that are extracted from the speech signal. The importance of variations in F_0 is known in the literature. However, in the presented work, we explored the dynamic variation in the *SoE1,* and *SoE2* features that are derived from speech and the excitation source, $\dot{g}(t)$, as well. In the case of natural human speech production mechanism, there is a movement of the vocal folds that actually affects the F_0 and the amplitude of speech. Such a representation is absent in computer generated speech (especially, vocoder-based speech). The source-based features on its own cannot significantly reduce the performance of SSD systems. However, fusion with the system-based features assisted in reducing the % EER for vocoder-based spoofing attacks. In addition, for the case of completely unknown attacks, the source-based features also showed a consistent decrease in the % EER with the dynamic information (for almost all the systems of the Blizzard dataset).

5.3 Prediction Techniques of Speech for SSD task

Historically, the idea of Linear Prediction (LP) and all-pole modeling of the system was used in system identification and control literature [**186**] and then it was brought to the field of speech signal processing [**185**], [**187**]. This Section presents the various prediction techniques that have been used in this thesis for the SSD task. The motivation behind using prediction analysis is that the speech signal can be considered to have two types of correlation, i.e., *short-term* and *long-term*. The short-term correlation occurs over the interval of the vocal tract impulse response within a pitch period (i.e., T_0) while the long-term correlation occurs across consecutive pitch periods [**21**]. The first work to explore prediction-based features explicitly for SSD was carried out in [**76**]. The basic motivation behind this was that the SS and VCS are quite likely to be either very easily predicted (if generated with a simplified acoustic model) or very difficult to predict (if artifacts are present in the signal as in the case of joints for USS-based speech). Hence, the combination of LP followed by LTP (i.e., LP-LTP) will first predict the short-term correlations and then

the long-term correlations. In the case of the natural speech signal, as the dependencies are more between the samples, both LP prediction error and LTP prediction error may be high as compared to that of spoofed speech. Hence, the energies along with their ratios and further derived features will be different for natural and spoofed speeches which can be used as countermeasures as in [**76**].

In this work, we propose using NLP in place of both LP and LTP to obtain better prediction. Using NLP instead of LTP at the second stage will also reduce the error. However, this is true for the short-term case and not for the long-term. On the other hand, using NLP in place of LP (i.e., before LTP) will reduce the NLP error and hence, the ratio between NLP and LTP residual energy will not be too diverse to have better discriminative properties. We study all these cases in the next sub-Section and discuss its suitability for the present task of detecting spoofed speech.

5.3.1 Linear Prediction (LP) of Speech

The speech signal is produced by the *convolution* of the excitation source and time-varying vocal tract system components. This excitation source and vocal tract system components are to be separated, in principle, from the speech signal to study them independently for several speech analysis applications such as pitch (F_0) detection and formant estimation, etc. Next, the LP analysis is discussed in brief [**188**]. For a speech signal $s(n)$ of length N, the predicted signal $\hat{s}(n)$ can be represented using p predictor memory element as follows,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \tag{5.7}$$

where a_k*'s* are the LP coefficients (LPC), and the prediction error $e_{LP}(n)$ or the LP residual can be computed (from eq. (5.7)) as,

$$e_{LP}(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{5.8}$$

It can be said that the prediction coefficients, $\{a_k\}$ where $k \in [1,p]$ are able to efficiently model the speech signal within a particular frame based on the prediction gain G_p and it is defined as [**76**],

$$G_p = \frac{E_x}{E_e}, \tag{5.9}$$

where E_x and E_e are the short-time energy (mean squared value) of original speech and predicted error signal, respectively. If G_p is high, the prediction of speech samples is better. The methods for obtaining the *optimal* predictor coefficients are based on minimizing the l^2 energy of LP residual signal and it is given by,

$$E = \sum_{n=0}^{N-1} | e_{LP}(n) |^2. \tag{5.10}$$

The LP analysis has been used extensively for estimating the pitch period or epochs or GCIs of the given speech given segment [**189**], formant frequency and bandwidth estimation [**190**], etc. It is extensively used for speech coding purposes both in time-domain [**191**] as well as in the frequency-domain [**192**].

5.3.2 Long-Term Prediction (LTP) of Speech

Due to nonlinear interaction of the source and the vocal tract system, the LP residual has high-intensity peaks (pitch pulses) at the GCIs. This is due to the nonlinear interaction of the source due to impulse-like excitation and the vocal tract system. The speech samples around the GCIs are a result of this nonlinear interaction which cannot be predicted by the LP technique. The basic idea of LTP is to remove the pitch pulses in the LP residual that are correlated and predictable over consecutive pitch periods, i.e.,

$$e_{LP}[n] \approx b \cdot e_{LP}[n - T_0], \tag{5.11}$$

where b is the scale factor and T_0 is the pitch period that can be calculated from the autocorrelation function of $s(n)$. The long-term predictor is of the form,

$$B(z) = 1 - bz^{-p}, \tag{5.12}$$

where bz^{-p} is the long-term predictor in the z-domain. The output of the long-term prediction error filter is,

$$e_{LTP}[n] = e_{LP}[n] - b \cdot e_{LP}[n - T_0]. \tag{5.13}$$

In LP analysis, the present sample is predicted based on the correlation of the p immediate past sample. On the other hand, in LTP, a current sample is predicted based on the correlation of a sample $s(n)$ at the n^{th} instant, with the similar samples which are a pitch period T_0 away from the sample $s(n)$ as shown in Figure 5.10 (Chap. 8 [**193**]). Thus, LTP operates on vectors rather than on individual samples. In

this case, a vector of samples can also be predicted using another vector of samples from the signal's history. The best matching vector is subtracted from LP residual error ($e_{LP}(n)$) resulting in LTP residual signal ($e_{LTP}(n)$). The LTP works efficiently with quasi-periodic voiced speech signals. The prediction error and prediction gain are calculated similarly as in the case of LP analysis.

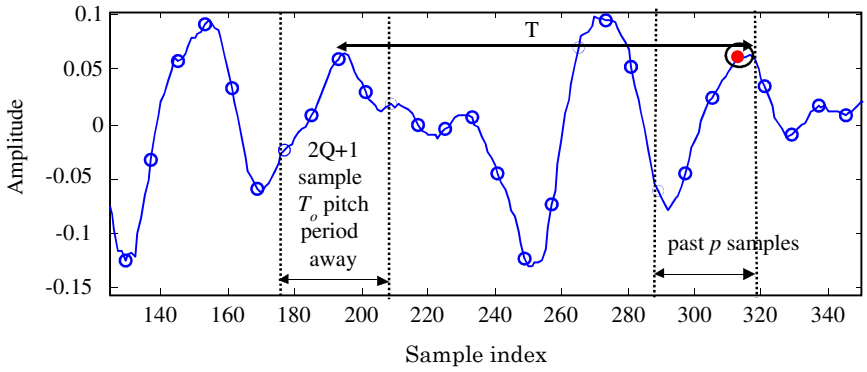


Figure 5.10: Schematic of the short-term correlation of a sample with *p* immediate past samples and the long-term correlation with the samples which are a pitch period *'To'* away. After [193**].**

5.3.2.1 Calculations of LTP Parameters

The parameter extraction procedure for LTP is explained as in [**194**]. Considering a speech signal, the LP residual (e_{LP}) can be obtained as per eq. (5.8). A particular frame size of *25* ms of the short-term LP residual signal (e_{LP} is denoted as d in this Section) is further processed in shorter sub-segments (of *5* ms, i.e., *40* samples for an *Fs=8* kHz). Therefore, if the e_{LP} frame is processed in m sub-segments, the sample related to the m^{th} sub-segment of the residual signal can be denoted as $d(k_m + k)$ with $m = 0,\ldots,3;\ k_m = k_0 + m*40$ and $k = 0,\ldots,39$ and k_0 denotes the first sample value of the current frame. In these calculations, for each sub-segments, a long-term correlation lag N_m, $(m = 0,\ldots 3)$, and gain factor is to be determined. These parameters for each sub-segment are implemented in following three steps [**194**].

- **Step 1:** Find the cross-correlation, $R_m(l)$, using current sub-segment of the short-term residual $d(k_m + i),\ (i = 0,\ldots,39)$ and previous samples of the reconstructed short-term residual signal $d'(k_m + i),\ (i = -120,\ldots,-1)$:

$$R_m(l) = \sum_{i=0}^{39} d(k_m + i) \cdot d'(k_m + i - l), \qquad \textbf{(5.14)}$$

where $m = 0,\ldots,3; k_m = k_0 + m*40$ and $l = 40,\ldots,120$. To find the above cross-correlation, consider the lags l greater than or equal to *40* and less than or equal to *120*, i.e., we can choose this l from the sample outside the current sub-segment and not delayed by more than two sub-segments.

- **Step 2:** Find the position N_m where the maximum peak of cross-correlation function occurs within this interval:

$$R_m(N_m) = max[R_m(l)], \qquad l = 40,\ldots,\ 120,\ \ m = 0,...,3 \tag{5.15}$$

- **Step 3:** Evaluate the gain factor, b_m, according to

$$b_m = \frac{R_m(N_m)}{S_m(N_m)}, \tag{5.16}$$

where $S_m(N_m) = \sum_{i=0}^{39} d'^2(k_m + i - N_m)$, for *m=0,...,3.*

In the above procedure, the short-term residual signal $d(k_0 + k)$ where $k = 0,...,159$ is processed by sub-segment of *40* samples. From each of the sub-segment of short-term residual samples (denoted here as $d(k_m + k)$), an estimate $d''(k_m + k),\ (k = 0,...,39)$ of the signal is subtracted to get the long-term residual signal $e_{LTP}(k_m + k),\ (k = 0,...,39)$,

$$e_{LTP}(k_m + k) = d(k_m + k) - d''(k_m + k). \tag{5.17}$$

The estimated samples $d''(k_m + k)$ are computed from the earlier reconstructed short-term residual samples d', adjusted to the current sub-segment LTP lag N_m' and weighted with the gain factor b_m' of the LTP sub-segment,

$$d''(k_m + k) = b_m' \cdot d'(k_m + k - N_m'), \tag{5.18}$$

where $m = 0,...,3; k = 0,...39; k_m = k_0 + m*40$. The LTP technique is widely used in speech coding (e.g., in GSM 06.10) or in narrowband and wideband adaptive multi-rate coders. Both LP and LTP analysis consider the linear combination of the speech samples which may be either short-term combination or long-term combination, respectively. However, in a speech signal, the sequences of samples are non-linearly correlated with each other. Hence, we study the NLP analysis of speech and explore its possible use in the SSD task.

5.3.3 Non-Linear Prediction (NLP) of Speech Signal

The set of Volterra functionals is *complete* (i.e., every Cauchy sequence converges to a *limit point* which belongs to the same *function* space). It means that every continuous functional of a signal $x(t)$ can be approximated with arbitrary precision as a sum of a finite number of Volterra functions in $x(t)$. This result was a generalization of the Weierstrass-Stone theorem (i.e., every continuous function of a variable $x(t)$ can be approximated with arbitrary precision as the sum of a finite number of polynomials in $x(t)$) [**195**]. To perform the nonlinear prediction of speech, consider a dynamical system with input data series $x(n)$ and the output $y(n)$ at time instant ($n=1,2,\ldots,N$), in multiples of sampling time τ. A power series expansion such as the Taylor series expansion may be used to describe the output of the system as,

$$y(t) = \sum\nolimits_{p=0}^{\infty} c_p x^p(t). \tag{5.19}$$

A nonlinear system with k memory terms can be then represented by means of an extension of eq. (5.7). This extension, known as the Volterra series expansion, which relates the input and output of the system is used. For a dynamical system, a closed-loop version of the Volterra series is used in which the output $y(n)$ feeds back as a delayed input (i.e., $x(n) \equiv y(n)$). Therefore, we analyze the univariate time series by using a discrete VW series of degree d and predictor memory k to calculate the predicted time series $\hat{y}(n)$ given by:

$$\hat{y}(n) = a_0 + a_1 y(n-1) + a_2 y(n-2) + ... + a_k y(n-k) + a_{k+1} y(n-1)^2 + a_{k+2} y(n-1) \times y(n-2) + a_{M-1} y(n-k)^d,$$

$$\therefore \hat{y}(n) = \sum\nolimits_{m=0}^{M-1} a_m z_m(n), \tag{5.20}$$

where the functional basis $\{z_m(n)\}$ is composed of all the distinct combinations of the *embedding space* coordinates up to degree d with a total dimension $M=(k+d)!/k!d!$ [**196**]. Thus, each model is parameterized by k and d corresponding to the *predictor memory* and the *degree* of nonlinearity in the model, respectively. The coefficients a_m'*s* in eq. (5.20) are estimated by Korenberg's fast algorithm using Gram-Schmidt procedure from the linear and nonlinear autocorrelation of the data-series itself [**196**]. From eq. (5.7) and eq. (5.20), for NLP model, $s(n)=y(n)$ and $\hat{s}(n)=\hat{y}(n)$. Therefore, NLP residual, e_{NLP} is given by [**197**],

$$e_{NLP}(n) = y(n) - \hat{y}(n). \tag{5.21}$$

The NLP has been used in several applications such as speaker recognition and estimation of epochs or GCIs from the speech signal [**198**]. The use of NLP in speech analysis-by-synthesis [**199**] and speech coding [**200**] also been explored.

5.3.4 Prediction Analysis of Natural and Spoofed Speech

The LP and NLP techniques use short-term dependencies in the past samples of a speech segment to predict the current sample. On the other hand, LTP uses long-term dependencies in the speech signal (i.e., past samples T_0 pitch period away from the current samples) to predict the speech sample. To show which of the three prediction techniques are efficient, we compare the LP, LTP, and NLP residuals as in Figure 5.11. It is observed that the NLP residual had less energy than LTP and LP residual. This was observed over several such voiced regions.

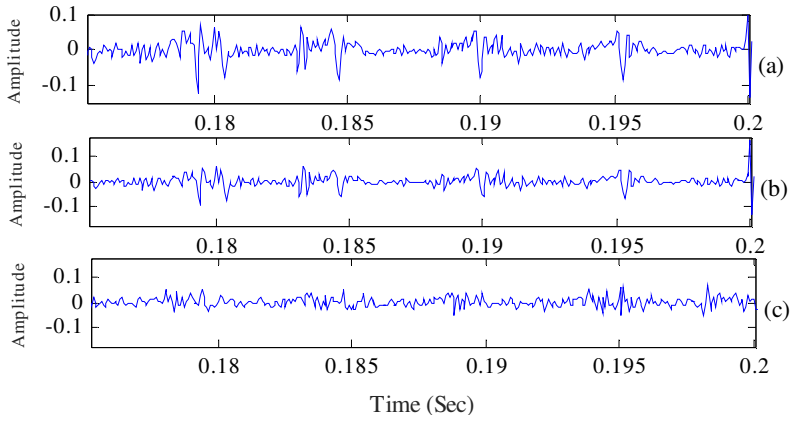


Figure 5.11: Comparison between LP, LTP and NLP (a) LP residual (b) LTP and (c) NLP residual for voiced speech.

To further statistically quantify it, the average l^2 norm is estimated for all residuals over *100* natural utterances of the *D1* speaker of the ASVspoof 2015 challenge database. For LP analysis d=1 and for NLP d=2 is considered. To keep the number of coefficients constant for LP and NLP analysis, p=*2*, *14*, *20* are considered and corresponding to k=*1*, *4*, *5*, giving a total of *3*, *15* and *21* coefficients, respectively. The average l^2 energy as in eq. (5.10) is shown in Table 5.7. From this result, we conclude that the energy of prediction error decreases for NLP as compared to LP and LTP for the same number of coefficients. Similar analysis between LP and NLP using the l^2 norm of residual signal over several utterances has been shown in [**197**].

Table 5.7: Average l^2 norm energy of LP, LTP and NLP residual signal over *100* utterances

Feature set	Total no. of coefficients in prediction		
	3	15	21
LP residual	12.3	6.42	6.18
LTP residual	**10.5**	5.52	5.26
NLP residual	**10.5**	**3.63**	**3.16**

5.3.5 Countermeasures for Spoofed Speech Detection

In this Section, we discuss the features that are presented in [**76**] based on LP-LTP combination and use them to propose the NLP-LTP and LP-LTP combination. All the three combinations are shown in Figure 5.12.

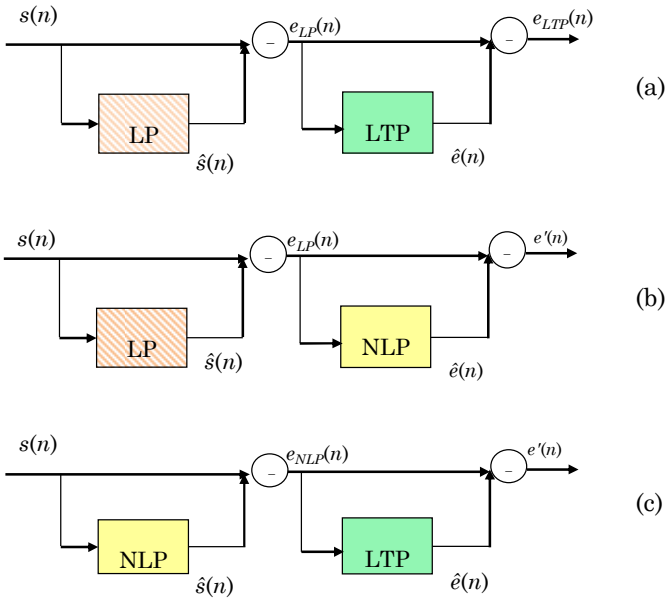


Figure 5.12: Schematic of the various prediction techniques combinations for detecting spoofed speech (a) LP-LTP (After [76]), (b) LP-NLP (After [90]) and (c) NLP-LTP.

5.3.5.1 LP-LTP Combination

The LP-LTP combination as proposed in [**76**] is shown in Figure 5.12 (a). The speech signal $s(n)$ is predicted using LP and thereafter, the LP residual is obtained as in eq. (5.8), i.e., the predicted values are subtracted from the original samples. The resulting LP residual is processed further to the next block of LTP that operates on vectors of samples rather than on individual samples. When the best matching vector is found, it is subtracted from the signal, resulting in LTP prediction error signal, e_{LTP}. The present LP-LTP approach is different from that in [**76**] in the sense that here we carry LP and LTP analysis at the frame-level. It is known that the LP

residual is intelligible on hearing, i.e., there is still sufficient dependency and correlation between the sequence of samples in the LP residual that is further captured by LTP. In addition, the LP residual error is known to have peaks at GCIs due to nonlinear interaction. The nonlinear interaction is due to the various aerodynamic aspect of speech production. Thus, the LP residual has primary pitch pulses that can be removed or minimized by LTP. Based on this architecture, features are proposed for spoof detections as described below [**76**]:

- m1: *MeanLPerr*: mean energy of the LP error, i.e., mean energy of e_{LP},
- m2: *MeanLTPerr*: mean energy of the LTP error, i.e., mean energy of e_{LTP},
- m3: *MaxLTPerr:* maximum energy of the LTP error, e_{LTP},
- m4: *MeanLTPgain:* mean LTP gain (i.e., mean ratio of energies of the LP and LTP errors, mean *Gp* as in eq. (5.9)),
- m5: *MaxLTPgain:* maximum of the *Gp* for LTP,
- m6: *MeanErrLen*: mean length of segments with e_{LTP} above threshold (θ),
- m7: *MaxErrLen*: maximum length of segments with e_{LTP} above θ,
- m8: *MeanNoErrLen*: mean length of segments with e_{LTP} equal or below θ,
- m9: *MaxNoErrLen*: maximum length of segments with e_{LTP} equal or below θ,
- m10: *EnergyLP*: total energy of LP residual, e_{LP},
- m11: *EnergyLTP*: total energy of LTP residual, e_{LTP}.
- m12: *ErrChangeRate:* LTP threshold crossing rate (counted per *20* ms frame),

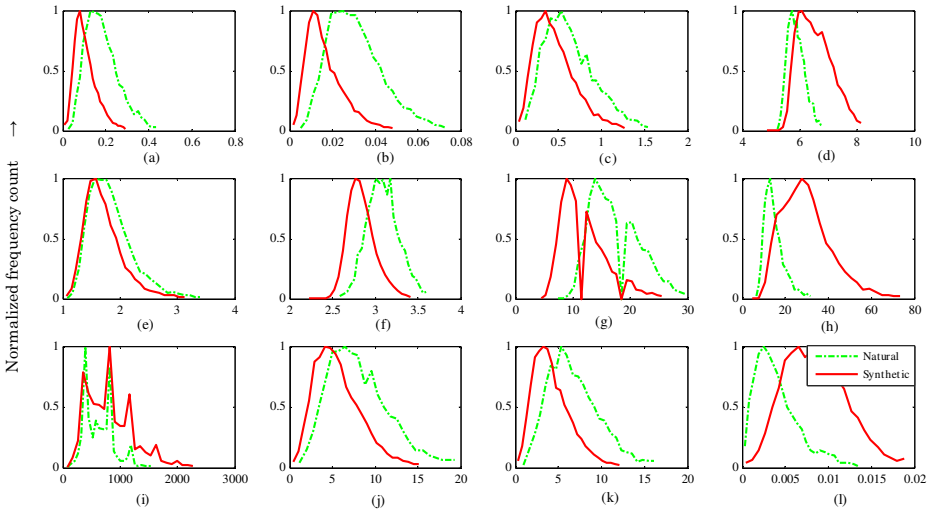


Figure 5.13: Histogram of the various spoofing countermeasures (m1-m12) of the LP-LTP-based combination (a) meanLPerr (b) MeanLTPerr (c) MaxLTPerr (d) MeanLTPgain (e) MaxLTPgain (f) MeanErrLen (g) MaxErrLen (h) MeanNoErrLen (i) MaxNoErrLen (j) EnergyLP (k) EnergyLTP and (l) ErrChangeRate.

As compared to the features in [**76**], features m10-m11 are added in the present work to include the total energy value than just mean energy. The distribution of LP-LTP, LP-NLP and NLP-LTP-based feature sets using *3750* natural utterances (green) and *12652* spoofing utterances (red) of the training set of ASVspoof challenge dataset is shown in Figure 5.13, Figure 5.14 and Figure 5.15, respectively. It is observed in Figure 5.13 that there exists less overlap between the histogram of natural speech and spoofed speech showing better discriminating properties. It is observed that distribution of features like MeanErrLen, MaxErrLen, MeanNoErrLen and ErrChangeRate is different for natural and spoofed speech.

5.3.5.2 LP-NLP Combination

The schematic of LP-NLP combination is shown in Figure 5.12 (b). This architecture is similar to LP-LTP in the sense that the long-term prediction which reduces the error of the resultant LP residual is replaced by NLP for which also the resultant error is less than LP analysis. The difference lies in the fact that LTP captures long-term linear dependencies in the speech signal while NLP captures short-term nonlinear dependencies. The NLP is carried on the LP residual rather than the speech signal to fit in the LTP framework as discussed in Section 5.3.5.1.

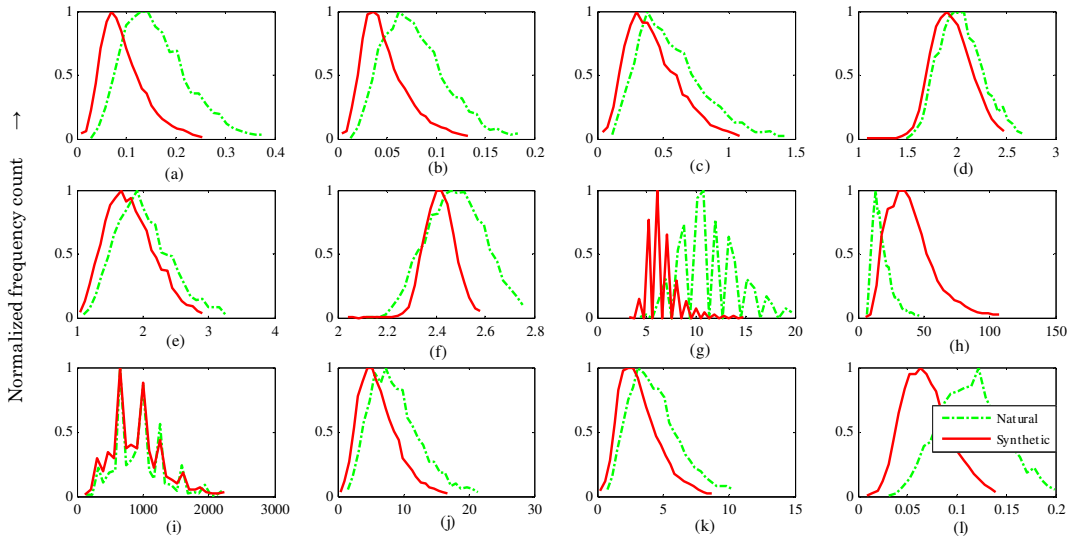


Figure 5.14: Histogram of the various spoofing countermeasures (m1-m12) of the LP-NLP-based combination (a) meanLPerr (b) MeanNLPerr (c) MaxNLPerr (d) MeanNLPgain (e) MaxNLPgain (f) MeanErrLen (g) MaxErrLen (h) MeanNoErrLen (i) MaxNoErrLen (j) EnergyLP (k) EnergyNLP and (l) ErrChangeRate.

The countermeasures proposed in the form of m1-m12 are similar except that the LTP is replaced by the NLP. These countermeasures are proposed in [**90**]. It is observed in Figure 5.14 that MaxErrLen, MeanNoErrLen and ErrChangeRate were more discriminating for natural and spoofed speeches than the other features. These features were similar to that of the LP-LTP combination.

5.3.5.3 NLP-LTP Combination

The schematic of NLP-LTP combination is shown in Figure 5.12 (c). This architecture is similar to LP-LTP in the sense that the long-term prediction is applied to the residual obtained by short-term prediction. However, the short-term prediction by NLP is better than LP as it captures the nonlinear dependencies in the speech samples. This may prove to be a drawback, as both NLP and LTP provide better prediction and hence, the ratios of energy or the features from this combination may not be significantly different for natural and spoofed speech. The countermeasures proposed in the form of m1-m12 are similar except that the LP residual is replaced by the NLP residual.

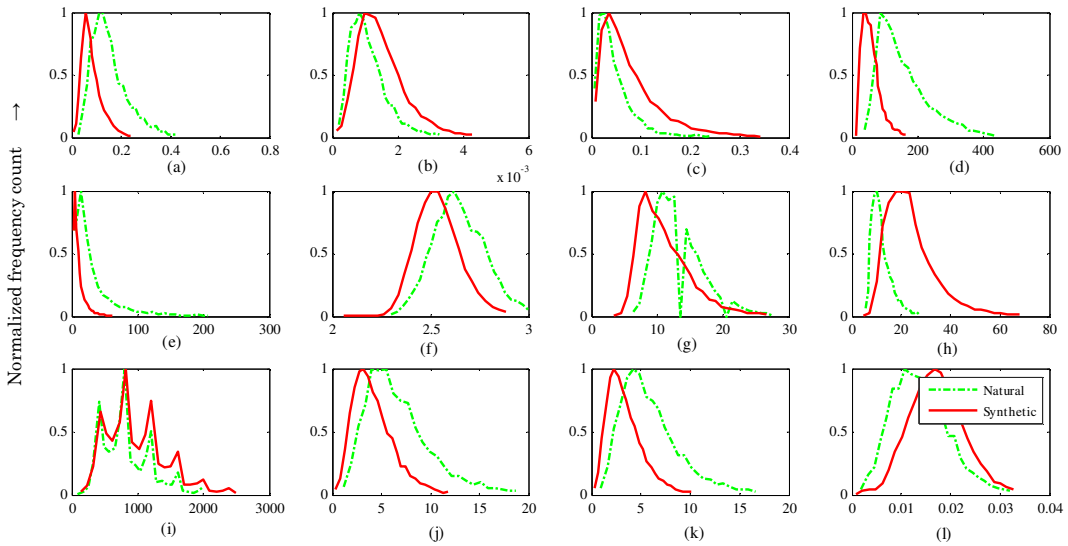


Figure 5.15: Histogram of the various spoofing countermeasures m1-m12 of the NLP-LTP-based combination (a) meanNLPerr (b) MeanLTPerr (c) MaxLTPerr (d) MeanLTPgain (e) MaxLTPgain (f) MeanErrLen (g) MaxErrLen (h) MeanNoErrLen (i) MaxNoErrLen (j) EnergyNLP (k) EnergyLTP and (l) ErrChangeRate.

Figure 5.15 shows that unlike the features from LP-LTP and LP-NLP combination, the NLP-LTP-based features do not show much discrimination between the natural and spoofed speech. Therefore, possibly the discrimination due to these features

might not show a significant reduction in the % EER. However, as shown in Figure 5.13 (d), the MeanLTPgain shows some discrimination, which is not the case of LP-LTP and LP-NLP combination. Therefore, the individual performance of the features might not be significant. However, a fusion of the NLP-LTP combination might add complementary information to the LP-LTP and LP-NLP combination.

5.3.6 Experimental Results

5.3.6.1 Parameterization

For LP analysis, the pole order $p=20$ is considered (due to the relationship between sampling frequency *Fs* and length of vocal tract [**185**]). For LP-LTP and NLP-LTP combination a frame length of *25 ms* is chosen to predict using LP and NLP. Then LP and NLP residual is calculated after subtracting predicted signal (obtained using the estimated coefficients) from the original speech signal. For LP-LTP and LP-NLP combination, the LP residual is framed at *5 ms* window. Furthermore, for NLP, $k=5$ and $d=2$ is considered in eq. (5.20). The LP-LTP, LP-NLP and NLP-LTP features form a *12-dimension (12-D)* feature vector of m1 to m12 for the entire speech signal. The entire analysis is done for voiced regions. For ErrChangeRate feature in all the combinations, the threshold value is set to $\theta=0.02$ [**76**]. For further representation, LP-LTP, LP-NLP and NLP-LTP feature sets are abbreviated as M1, M2 and M3, respectively. In [**76**], the features are extracted by processing sample-by-sample on the speech signal whereas in our work we consider the traditional frame-level processing for faster computation and to facilitate comparison with frame-based system features that are discussed in Chapter 4.

5.3.6.2 Results on the Development Set of ASVspoof challenge Database

The results in EER on the development set for the M1, M2, and M3 feature sets are shown in Table 5.8. As shown in Table 5.8, the % EER of M1, M2 and M3 on development set are *4.77*, *9.17* and *13.89*, respectively. As discussed in Section 5.3.5.3, due to better prediction by both NLP and LTP, this combination (i.e., M3) gives high % EER. It is observed that the best score-level fusion of the features with each other occurs at $\alpha_f = 0.3$.

Fusion with system-based features: The best fusion of M1 and M2, M1 and M3 and M2 and M3 at α_f =*0.*3 are then fused at score-level with the *36-D* (*12*s+*12*-Δ+*12*-ΔΔ) system-based MFCC, CFCC, CFCCIFS and SBAE feature sets as in Table 5.9.

Table 5.8: EER (in %) for M1, M2 and M3 feature set along with their score-level fusion at various fusion factor α_f on the development set

Feature Set 1	Fusion Factor (α_f)											Feature Set 2
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
M1	4.77	4.14	3.71	**3.60**	3.60	3.83	4.28	4.97	5.77	7.32	9.17	**M2**
M1	4.77	4.26	3.91	**3.88**	4.17	4.8	5.83	7.23	9.09	11.32	13.89	**M3**
M2	9.17	8.6	8.12	**7.94**	8.20	8.57	9.17	9.92	11	12.43	13.89	**M3**

Score-level fusion is carried as per eq. (3.6)

Table 5.9: EER (in %) for score-level fusion of best combination of M1-M2, M1-M3 and M2-M3 with system-based feature sets (using *D3* feature vector) at various fusion factors α_f on the development set

Feature Set 1	Fusion Factor (α_f)											Feature Set 2
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
best (M1-M2)	3.60	3.32	3.03	2.77	2.46	2.14	1.72	1.34	0.92	**0.57**	1.60	**MFCC**
best (M1-M2)	3.60	3.37	3.06	2.77	2.52	2.17	1.92	1.34	1.03	**0.66**	1.54	**CFCC**
best (M1-M2)	3.60	3.40	3.15	2.89	2.63	2.43	1.94	1.49	1.06	**0.51**	1.23	**CFCCIFS**
best (M1-M2)	3.60	3.40	3.17	2.89	2.66	2.37	2.03	1.60	1.14	**0.57**	1.49	**SBAE**
best (M1-M3)	3.89	3.63	3.29	3.06	2.66	2.32	1.89	1.46	1.03	**0.63**	1.60	**MFCC**
best (M1-M3)	3.89	3.63	3.37	3.09	2.77	2.37	2.03	1.54	1.06	**0.69**	1.54	**CFCC**
best (M1-M3)	3.89	3.72	3.43	3.23	2.97	2.60	2.17	1.69	1.12	**0.57**	1.23	**CFCCIFS**
best (M1-M3)	3.89	3.72	3.46	3.17	2.92	2.63	2.23	1.72	1.17	**0.69**	1.49	**SBAE**
best (M2-M3)	7.95	7.66	7.21	6.66	5.83	5.12	4.23	3.26	2.23	**1.26**	1.60	**MFCC**
best (M2-M3)	7.95	7.64	7.15	6.69	6.06	5.12	4.32	3.35	2.46	**1.37**	1.54	**CFCC**
best (M2-M3)	7.95	7.72	7.43	6.95	6.35	5.69	4.72	3.69	2.63	**1.26**	1.23	**CFCCIFS**
best (M2-M3)	7.95	7.66	7.26	6.81	6.23	5.55	4.63	3.75	2.46	**1.29**	1.49	**SBAE**

Score-level fusion is carried as per eq. (3.6)

It is observed from Table 5.9 that at α_f = *0.9*, i.e., *10* % of excitation source-based features and *90* % of system-based features, the performance of system-based features is reduced even more than *50* % on using the best combination of M1-M2 features and best combination of M1-M3 features. On the other hand, the decrease in % EER of system-based features with the M2-M3 combination is relatively less. The least EER is *0.51* % obtained by fusion of best (M1-M2) with CFCCIFS feature set.

Dependency on spoofing algorithms: Figure 5.16 shows the spoof dependency of the prediction-based features. It is observed that for the known attacks, the M1 feature set gave the least % EER as compared to M2 and M3 that could not identify the known type that well. For the same type of attacks, considering VCS, the % EER

is least for M2 combination indicating that the linear and nonlinear combination is useful for spoof detection. Hence, for VCS, it may be possible that the nonlinearity might be absent during the speech generation process. The SS spoof could identify its similar type as good as the known case. Next, for the different type of attacks, for VCS spoof, the detection was much better from M1 to M3. However, on training with the SS speech, M2 and M3 did not predict the different type of spoof (i.e., VCS) as well as with the M1 combination. Thus, for the known case and the same type of attacks, the M1 and M2 techniques perform well respectively. However, for the different type of attacks, the performance of the features depends on the type of spoofed speech (SS or VCS) that is used in training.

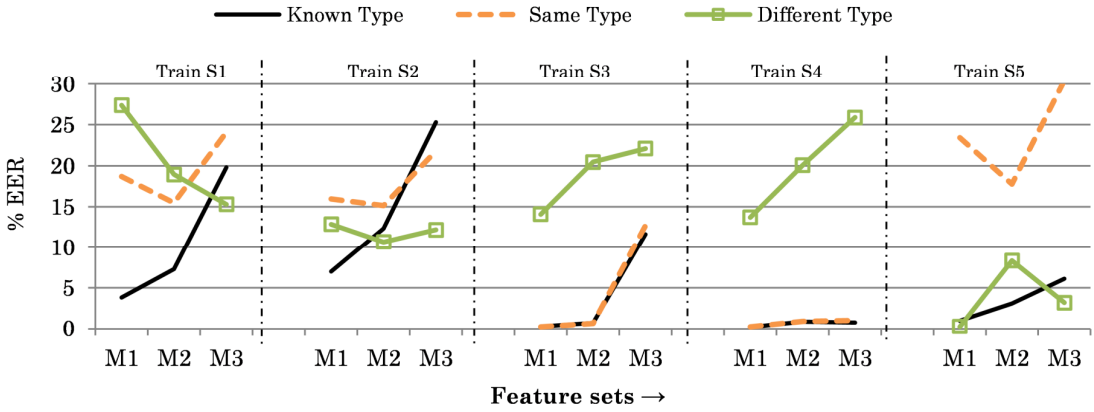


Figure 5.16: The % EER on known, same and different type of attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for prediction-based M1, M2 and M3 feature sets and tested on the development dataset.

Discussion on the DET curves: The DET curves for the M1, M2 and M3 feature sets and their score-level fusion at $\alpha_f = 0.3$ are shown in Figure 5.17(a).

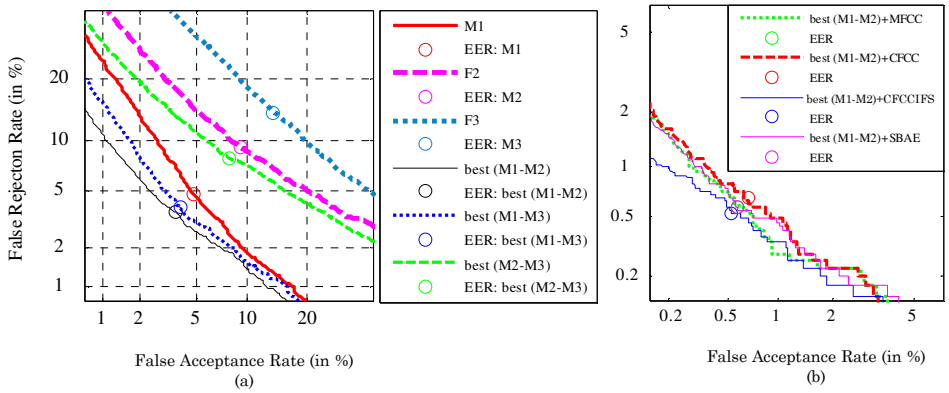


Figure 5.17: DET curves on the development set for (a) M1, M2 and M3 feature sets and their best score-level fusion and (b) the best score-level fusion of M1 and M2 with the system-based features.

It is observed that the best % EER is due to the M1 and M2 combination (i.e., best (M1-M2)), which is better than all the feature sets along the entire DET curve. The performance of FRR for a particular value of FAR is better after fusion with M2 feature set. The DET curve for score-level fusion of best (M1-M2) with the system-based features at α_f=*0.9* is shown in Figure 5.17(b). The improvement in both system-based features and the best combination of M1-M2 after their score-level fusion is observed. The DET curve for M1 and M2 combination after fusion with CFCCIFS shows the best performance in terms of % EER and also along FRR.

5.3.6.3 Results on the Evaluation Set of ASVspoof challenge Database

The results on the evaluation set are reported in Table 5.10. The M1, M2 and M3 feature sets report an EER of *10.91* %, *14.14* % and *18.73* %, respectively. The fusion of the M1, M2 and M3 features among itself shows that the best achieved EER is *9.73* % with the fusion of M1 and M2. The other combinations did not perform better than this combination. On the development set, the best fusion factor for score-level fusion of M1 and M2, M1 and M3, and M2 and M3 obtained was α_f=*0.3*, which on the evaluation set was observed to be *0.4*, *0.3* and *0.2*, respectively. This slight change in

Table 5.10: EER (in %) for score-level fusion of best combination of M1-M2, M1-M3 and M2-M3 with system-based feature sets (using *D3* feature vector) at various fusion factors α_f on the evaluation set

Feature Set 1	Fusion Factor (α_f)											Feature Set 2
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
M1	10.91	10.44	10.05	9.83	**9.73**	9.84	10.12	10.65	11.47	12.65	14.14	**M2**
M1	10.91	10.65	10.49	**10.45**	10.54	10.97	11.67	12.88	14.47	16.51	18.73	**M3**
M2	14.14	13.68	**13.41**	13.83	13.53	13.89	14.44	15.14	16.16	17.38	18.73	**M3**
best (M1-M2)	9.73	9.52	9.32	9.05	8.76	8.4	8.03	7.53	6.62	5.12	**4.26**	**MFCC**
best (M1-M2)	9.73	8.86	8.25	7.74	7.28	6.80	6.16	5.36	4.13	2.41	**1.74**	**CFCC**
best (M1-M2)	9.73	9.16	8.62	8.10	7.63	7.09	6.49	5.68	4.63	2.87	**1.60**	**CFCCIFS**
best (M1-M2)	9.73	9.54	9.33	9.07	8.77	8.39	7.84	7.06	5.70	3.56	**2.49**	**SBAE**
best (M1-M3)	10.45	10.24	9.96	9.65	9.3	8.91	8.41	7.81	6.87	5.38	**4.26**	**MFCC**
best (M1-M3)	10.45	9.57	8.86	8.31	7.80	7.20	6.51	5.67	4.34	2.49	**1.74**	**CFCC**
best (M1-M3)	10.45	9.90	9.27	8.72	8.15	7.56	6.91	6.04	4.82	2.90	**1.60**	**CFCCIFS**
best (M1-M3)	10.45	10.26	10.01	9.72	9.37	8.86	8.21	7.38	6.01	3.81	**2.49**	**SBAE**
best (M2-M3)	13.41	12.99	12.53	11.98	11.34	10.52	9.57	8.44	7.13	5.41	**4.26**	**MFCC**
best (M2-M3)	13.41	12.53	11.78	11.03	10.19	9.18	8.09	6.72	5.05	2.88	**1.74**	**CFCC**
best (M2-M3)	13.41	12.83	12.19	11.52	10.76	9.91	8.87	7.53	5.72	3.37	**1.60**	**CFCCIFS**
best (M2-M3)	13.41	13.04	12.61	12.11	11.52	10.71	9.69	8.29	6.45	3.88	**2.49**	**SBAE**

Score-level fusion is carried as per eq. (3.6)

the fusion factors is due to the presence of unknown attacks in the evaluation data set. The results of score-level fusion of the best fusion of M1-M2, M1-M3, M2-M3 with the system-based features at various α_f is shown in Table 5.10. It is observed that even after fusion with the prediction-based features, the system-based features do not show any improvement in the % EER. These results are not consistent with the results obtained in the development set, where a significant improvement was observed over the system-based features. The reason behind this is that the evaluation set consists of both vocoder-dependent and vocoder independent *S10* spoof (which is made by concatenating speech sound units directly). Hence, it is necessary to observe the results in % EER with and without *S10* spoof.

The best fusion factors for M1-M2, M1-M3 and M2-M3 combination are *0.4*, *0.3* and *0.2*, respectively (as shown in Table 5.10). These best combination of source features are fused at α_f = *0.9* with the system-based features. Using these fusion factors, the % EER for only vocoder-based speech (without *S10*) and with *S10* spoof on the evaluation set is shown in Table 5.11.

Table 5.11: EER (in %) in terms of individual attacks, average known attacks, average unknown attacks, average with and without *S10* spoof for M1, M2 and M3 feature sets, their best fusion combination and score-level fusion with the system-based feature sets (using *D3* feature vector) at selected α_f on the evaluation set

Feature Sets	Individual Attacks										Average			
	Known Attacks					Unknown Attacks								
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn.	Ukn.	w/o S10	Avg.
M1	5.74	4.13	0.02	0.02	1.45	2.55	7.49	0.04	0.95	86.66	2.27	19.54	2.49	10.91
M2	9.07	7.70	1.67	1.45	4.55	10.68	24.50	0.97	3.27	77.59	4.89	23.40	7.09	14.14
M3	19.33	18.29	0.64	0.73	5.68	26.70	40.15	0.39	2.42	72.63	8.93	28.46	12.70	18.70
best(M1-M2)	2.67	1.82	0.00	0.00	0.48	0.97	6.99	0.02	0.48	83.87	0.99	18.47	1.49	9.73
best(M1-M3)	4.09	2.98	0.01	0.01	0.63	1.77	9.57	0.01	0.47	84.83	1.54	19.33	2.17	10.44
best(M2-M3)	7.68	6.45	0.55	0.50	3.02	10.16	25.33	0.42	1.75	78.01	3.64	23.13	6.21	13.39
best(M1-M2)+ MFCC	0.00	0.04	0.00	0.00	0.02	0.02	0.01	0.00	0.01	51.11	0.01	10.23	**0.01**	5.12
best(M1-M3)+ MFCC	0.00	0.05	0.01	0.01	0.04	0.06	0.05	0.00	0.01	53.48	0.02	10.72	**0.03**	5.37
best(M2-M3)+ MFCC	0.00	0.20	0.00	0.00	0.11	0.26	0.11	0.00	0.01	53.37	0.06	10.75	**0.08**	5.40
best(M1-M2)+ CFCC	0.01	0.16	0.00	0.00	0.13	0.11	0.15	0.00	0.05	23.52	0.06	4.77	0.07	**2.41**
best(M1-M3)+ CFCC	0.01	0.26	0.01	0.01	0.10	0.16	0.25	0.00	0.03	24.02	0.08	4.89	0.09	**2.48**
best(M2-M3)+ CFCC	0.05	0.57	0.00	0.00	0.52	0.70	0.54	0.02	0.05	26.32	0.23	5.52	0.27	**2.88**
best(M1-M2)+ CFCCIFS	0.02	0.08	0.00	0.00	0.07	0.07	0.19	0.00	0.02	28.22	0.03	5.70	0.05	2.87
best(M1-M3)+ CFCCIFS	0.02	0.14	0.01	0.01	0.05	0.11	0.34	0.00	0.02	28.29	0.04	5.75	0.08	2.90
best(M2-M3)+ CFCCIFS	0.03	0.33	0.00	0.00	0.36	0.59	1.17	0.09	0.05	31.03	0.14	6.59	0.29	3.36
best(M1-M2)+ SBAE	0.01	0.18	0.00	0.00	0.08	0.17	0.14	0.00	0.05	35.03	0.06	7.08	0.07	3.57
best(M1-M3)+ SBAE	0.02	0.32	0.01	0.01	0.07	0.27	0.28	0.00	0.05	36.98	0.08	7.52	0.11	3.80
best(M2-M3)+ SBAE	0.02	0.49	0.00	0.00	0.27	0.72	0.46	0.00	0.10	36.68	0.16	7.59	0.23	3.88

Score-level fusion is carried as per eq. (3.6), Kn=known, Ukn= Unknown, w/o S10=Average without S10, Avg. = Average of S1-S10

As reported in Chapter 4, the % EER of known attacks using MFCC, CFCC, CFCCIFS and SBAE features is *0.4* %, *0.8* %, *0.5* % and *1.0* %. After score-level fusion with best combination of M1 and M2, the % EER of system-based features decreases to *0.01 %*, *0.06* %, *0.03* % and *0.06* %, respectively. Likewise the average EER without *S10* was *0.31 %*, *0.56 %*, *0.46* % and *1.08* %, respectively, for MFCC, CFCC, CFCCIFS and SBAE features which decreases to *0.01* %, *0.07* %, *0.05* % and *0.07* %, respectively, on using the score-level fusion of best combination of M1 and M2 with the system-based features. The MFCC features performed best for known attacks and hence, in this case, the fusion with MFCC serves as a better distinguishing feature for vocoder-based spoofs. This performance of the score-level fusion of prediction-based features and system-based features turns out to perform better than the F_0, *SoE1* and *SoE2* features along with the dynamics (as discussed in Section 5.2.6.3) for the vocoder-based spoofs. For *S10* spoof, the % EER of the M1, M2 and M3 features used individually or by their best combination was found to be very high (~ *70* % to *80* %). Even after the fusion with system-based features, the % EER of *S10* spoof does not decrease, therefore, the average % EER turns out to be very high.

Dependency on spoofing algorithms: Figure 5.18 shows the spoof dependency of the prediction-based features on the evaluation set. As in several earlier examples, on the evaluation set, we show the analyses of the same type, different type and *S10* attack. The analyses of the known type are similar to that obtained on the development set and hence, not shown here. For the same type of attacks, the % EER is almost same for SS spoof using any of the M1, M2 or M3 combination of the feature set. On the other hand, for VCS spoof, its same type, i.e., VCS can be detected best using the M1 combination. The % EER increases with M2 and M3 feature set signifying that the same type of spoof detection by VCS training is well modeled using the linear and long-term prediction. For different type of attacks, for VCS spoof, the detection of SS spoof is much better using the NLP information embedded in the M2 and M3 feature set. The EER while detecting the SS spoof is *11.5* %, *7.5* % and *2.03* % on training with *S1*, *S2* and *S5* spoof, respectively, using M3 combination. For the SS speech, the detection of VCS was not well using the NLP-based information and hence, the % EER increased for the M2 and M3 feature set. Considering *S10* spoof separately, the % EER on training with any of the

available spoofs (SS or VCS) decreases with the information about the nonlinear and long-term prediction aspects of the speech signal. It was observed that with the use of NLP and LTP combination (i.e., M3), lower EER is obtained as compared to M1 and M2. The best EER obtained for the *S10* spoof is ~ *65* % which is very high and hence, the prediction-based features may not be effective to detect these spoofs.

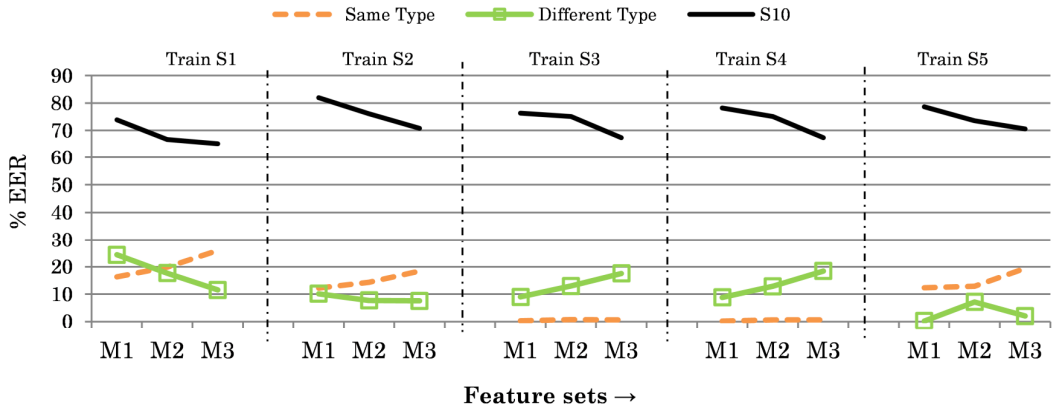


Figure 5.18: The % EER for the same type, different type and *S10* attack when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for prediction-based M1, M2 and M3 feature sets and tested on the evaluation dataset.

5.3.6.4 Results on the Blizzard Challenge 2012 Database

The results for the Blizzard Challenge 2012 database for the prediction-based M1, M2 and M3 features are shown in Table 5.12. The results on this database are contrary to that obtained on the ASV spoof challenge data. That is, the performance of M1 features is not better than that of M2 or M3. Unlike the F_0, *SoE1* and *SoE2* features, the prediction-based features did not generalize even amongst the unit-selection and statistical-based synthesis techniques.

5.3.6.5 Results on the Blizzard Challenge 2014 Database

The results for the Blizzard Challenge 2014 database for the prediction-based M1, M2 and M3 features are shown in Table 5.13. For the Gujarati dataset, the M3 features performed significantly well than M1 and M2 for all systems. For Hindi systems, HMM-based systems were identified well by M1 features and USS system by M3 features. Thus, the features do not generalize well for detecting unknown spoof which can be modified and explored further.

Table 5.12: EER (in %) for M1, M2 and M3 feature sets on training with the ASV spoof data and testing with the Blizzard Challenge 2012 database

Blizzard 2012	Systems	Feature Sets		
		M1	M2	M3
USS	B	56	**53**	54
Hybrid	C	54	49	**46**
Hybrid	D*	55	**37**	67
HMM	E*	10	41	**5**
USS	F	63	46	81
USS	G	73	**39**	84
HMM	H	42	42	**18**
USS	I	49	28	**6**
Diphone	J*	**27**	45	34
HMM	K*	43	51	**42**

* systems with lower MOS from *1* ≤ *2*

Table 5.13: EER (in %) for M1, M2 and M3 feature sets on training with the ASV spoof data and testing with the Blizzard Challenge 2014 database

Blizzard 2014	Gujarati systems	Feature Sets			Blizzard 2014	Hindi systems	Features Sets		
		M1	M2	M3			M1	M2	M3
HMM	C	21	63	**2**	HMM	B*	**13**	23	54
HMM	D	13	45	**2**	HMM	C	**18**	28	65
HMM	E	34	59	**2**	Hybrid	D	66	**50**	82
HMM-DNN	F	42	50	**7**	HMM	E	62	**33**	61
USS	G	29	68	**2**	HMM-DNN	F	62	36	**23**
HMM	H	47	55	**13**	USS	G	38	39	**29**
					HMM	H*	**10**	**10**	57
					HMM	K	**18**	32	77

* systems with lower MOS from *1* ≤ *2* (wavefiles for Gujarati system B and system I and Hindi system I are not available)

5.3.7 Summary

This Section discusses the various prediction schemes that can be used for the SSD task. The basic idea is that the natural speech has nonlinear dependencies between the sequences of samples. Hence, the use of nonlinear prediction techniques will aid in the detection of spoofed speech. On the ASV spoof data, the linear and long-term prediction features were found to perform better. In the case of the spoof dependency or the channel mismatch cases, where the training is done on one type of data and testing is done on the other type of speech, the nonlinear prediction features were found to have complementary information as compared to the linear and the long-term prediction features. However, the % EER was very high and the results were not consistent on the Blizzard datasets. These features could be modified for being used as more generalized countermeasures.

5.4 The Fujisaki Model

The *Fujisaki model* also known as the *command-response* model, is a prosodic model that represents F_0 contour in terms of the *phrase* and *accent* parameters as a result of the *translation* and *rotation* motion of the cricoids muscles, respectively [**201**]. The F_0 contour is known to be the result of movements of *intrinsic* muscles in the larynx [**202**]- [**203**] and Fujisaki model represents these movements. It is known that the F_0 carries linguistic and non-linguistic information and hence, the parameters of the Fujisaki Model that represents the movement of the muscles will also represent this information [**202**]. Here, we study variations in F_0 contour (in terms of Fujisaki model parameters) of natural and synthesized speech to study the source-based discriminative features between the two speech recordings. In the case of mimicry, an imitator tries to vary his or her F_0 contour so that the shape of the F_0 contour matches with the target speaker's F_0 contour [**14**]. However, speech synthesis technologies are not so human-like to perform such close matching to the F_0 contour of the natural speech signal. In other words, the natural speech is uttered with appropriate *breaks* and *accent* variations, which is not the case for synthetic speech. Therefore, we investigate the Fujisaki model parameters of natural and synthetic speech for discriminating these two speech recordings.

5.4.1 Physiological Interpretation

The movement of the thyroid cartilage relative to the cricoid cartilage has two degrees of freedom, namely,

- Horizontal movement due to *pars oblique* of the cricothyroid (CT) muscles and,
- Rotational movement due to activity of *pars recta* around the cricothyroid joint

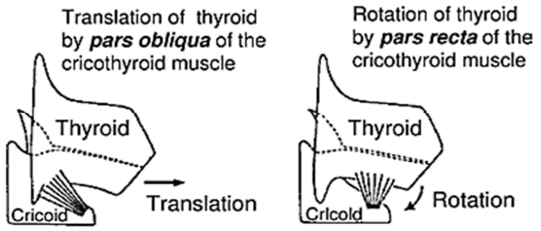


Figure 5.19: The role of pars obliqua and pars recta of the cricothyroid muscle in translating and rotating the thyroid cartilage. Adapted from [201**].**

The movement of the pars oblique and pars recta is shown in Figure 5.19. Therefore, an instant activity of the *pars oblique* of the CT contributing to the thyroid translation causes a change in $x_1(t)$ while the sudden increase or decrease in the activity of the *pars recta* of CT, contributing to the thyroid rotation causes a change in $x_2(t)$. Due to these two movements, the Fujisaki model is considered as a *superposition* of the *phrase component* (y_p) and *accent component* (y_a). The two contributions are superimposed with a constant value F_b (i.e., the minimum value of the speaker's F_0, which is known to be speaker-specific) to give a particular model generated F_0 contour. The various properties of Fujisaki model make it suitable for applications such as, F_0 contour analysis of various languages [**203**], language identification [**204**] and most importantly as a prosody model in speech synthesis [**23**].

5.4.2 Stress-Strain Relationship of Skeletal Muscles

It is shown that there exists a good *linear* relationship between the *tension* (stress) and the *stiffness* of the human vocal muscle, represented by the following differential equation [**201**],

$$\frac{dT}{dl} = a + bT, \tag{5.22}$$

where T is the tension, l indicates muscle length, and a is the stiffness at $T=0$. On solving eq. (5.22) with an integrating factor of $I(l) = e^{-bl}$, we get,

$$e^{-bl}T = \frac{-a}{b}e^{-bl} + c. \tag{5.23}$$

To estimate the constant c, substitute $l=l_o$ and $T=T_o$, therefore,

$$c = e^{-bl_o}T_o + \frac{a}{b}e^{-bl_o} = (T_o + \frac{a}{b})e^{-bl_o}. \tag{5.24}$$

Now, substituting c in the general solution eq. (5.24), we get,

$$T = (T_o + \frac{a}{b})e^{b(l-l_o)} - \frac{a}{b}, \tag{5.25}$$

where T_0 indicates the static tension applied to the vocal folds. When $T_0 >> a/b$, eq. (5.25) can be approximated by,

$$T = T_o e^{b(l-l_o)} = T_o e^{bx} = T_o \exp(bx), \tag{5.26}$$

where x indicates a change in vocal fold length from l_0 to l, when T_0 changes to T. The fundamental frequency (F_0) of vibration of an elastic membrane is given by,

$$F_o = c_o\sqrt{\frac{T}{\rho}}, \tag{5.27}$$

where ρ is the density per unit area of the membrane and c_0 is a constant inversely proportional to the *size* of the membrane, $\therefore \ \frac{F_o}{c_o} = \sqrt{\frac{T}{\rho}} \Rightarrow \quad F_o = c_o\sqrt{\frac{T_o}{\rho}}e^{bx/2}$.

Taking log (.) on both sides, $\log(F_o) = \log(c_o\sqrt{\frac{T_o}{\rho}}) + \frac{bx}{2}$,

$$\therefore \ \log(F_o) = \log(F_b) + \frac{bx}{2}. \tag{5.28}$$

The constant term written as F_b indicates the existence of baseline value of F_0 contour (i.e., the *minimum* value of the speaker's F_0). F_b is constant as long as the speaker has same speaking style and emotional state [**203**]. A time-varying component is added to F_b in log (F_0)-domain. Thus, eq. (5.28) shows log (F_0) changes only with x, i.e., log (F_0) changes with the change in length of the muscles ($x=l-l_o$). It is suggested that the change in length of the muscles have two movements. In particular, the horizontal *translation* due to the activity of *pars oblique* of the CT muscle and *rotation* around the CT joint due to *pars recta* of the CT muscle [**201**]-[**202**], i.e.,

$$\log(F_o) = \log(F_b) + \frac{b}{2}\{x_1(t) + x_2(t)\}. \tag{5.29}$$

Phrase Component: The phrase component models the pitch baseline, accounts for phrase wise slow overall declination line and it is characterized by a *fast rise* followed by a *slower fall*. The input to this model (*xp*) is composed of Dirac impulses, namely, *phrase commands*, located at the onsets of phrase activities. The *phrase control mechanism* is characterized by,

$$h_p(t) = \alpha^2 te^{-\alpha t}u(t), \tag{5.30}$$

where α [2;4] s^{-1} is its natural angular frequency and $h_p(t)$ is the impulse response of the phrase control mechanism as shown in Figure 5.20.

Accent Component: It models smaller-scale prosodic variations and accounts for accent variations. The input (*xa*) to this model is composed of rectangular pulses, namely, *accent commands*. The *accent control mechanisms*, are characterized by,

$$g_a(t) = [1 - (1 + \beta\, t)]e^{-\beta t}]u(t), \qquad \textbf{5.31}$$

where β[19;21] s^{-1} is its natural angular frequency and $g_a(t)$ is the step response of the accent control mechanism as shown in Figure 5.21.

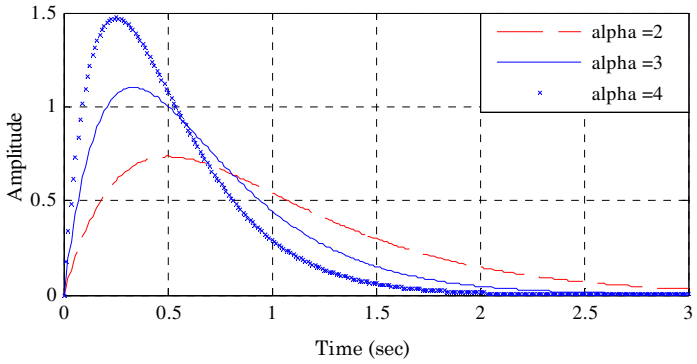


Figure 5.20: Impulse response of the phrase control mechanism.

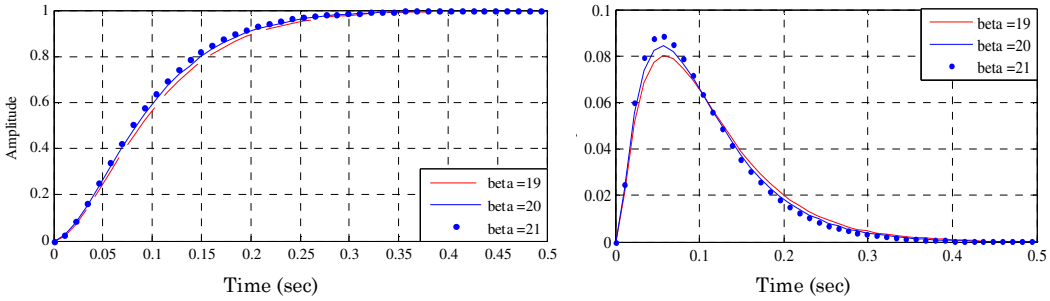


Figure 5.21: Step response (left) and the impulse response (right) of the accent control mechanism.

After obtaining the phrase and accent components, the entire F_0 contour of the utterance in the log-domain as in eq. (5.29) can be expressed as,

$$\log(F_0(t)) = \log(F_b) + y_p(t) + y_a(t), \qquad \textbf{(5.32)}$$

$$y(t) = y_p(t) + y_a(t) \;= \sum_{k=1}^{Np} A_{p,k} h_p(t - t_{p,k}) + \sum_{k=1}^{Na} A_{a,k}[g_a(t - t'_{a,k}) - g_a(t - t''_{a,k})], \qquad \textbf{(5.33)}$$

where N_p and N_a are the number of phrase and accent events; $A_{p,k}$ and $t_{p,k}$ are the magnitude and timing of the k^{th} phrase command; $A_{a,k}$, $t'_{a,k}$ and $t''_{a,k}$ are the magnitude, onset time and the end time of k^{th} accent command, respectively. The complete representation of the Fujisaki model is shown in Figure 5.22 where the nonlinear system for glottal airflow effects has been ignored [**202**].

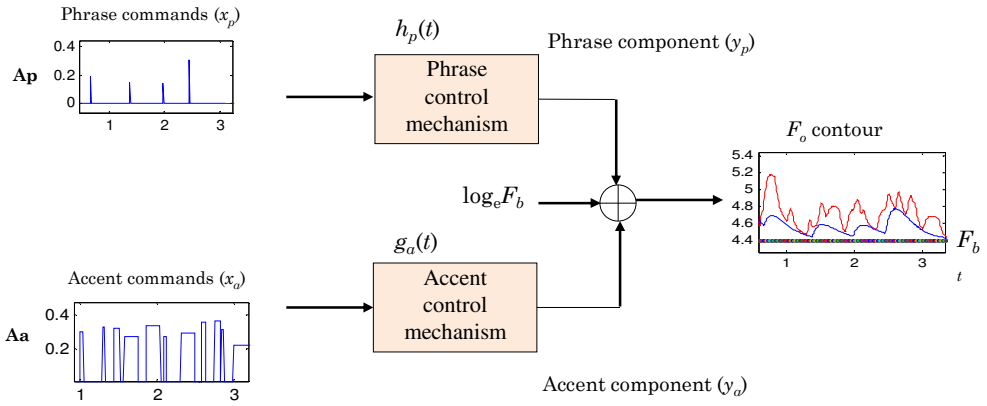


Figure 5.22: The Fujisaki model or functional command response model for generating *F0* contour. After [201], [202].

5.4.3 Extraction of Model Parameters

5.4.3.1 F0 Extraction

Extracting *prosodic* events from speech requires estimating *F0* contour of speech. As discussed in Section 5.2.3, the ZF method is used to estimate the *F0* contour [**182**]. The epoch locations, i.e., the GCIs are obtained from the negative-to-positive zero-crossings of the ZF filtered signal. Thereafter, the *F0* contour is obtained from the GCI locations. Fujisaki model requires a continuous contour and it deals with *macroprosody* only. Hence, two tasks are performed before modeling the *F0* contour, namely, intermediate values for unvoiced speech regions and short pauses are *interpolated* in the *F0* contour and microprosodic variations due to individual speech sounds units (such as plosive, fricatives, etc.) are smoothed out. Here, a linear fit is used during interpolation and then the *F0* contour has been smoothed prior to parameter estimation. Figure 5.23 shows the *F0* contour extracted from ZF filtering algorithm where the unvoiced regions are interpolated with a linear fit.

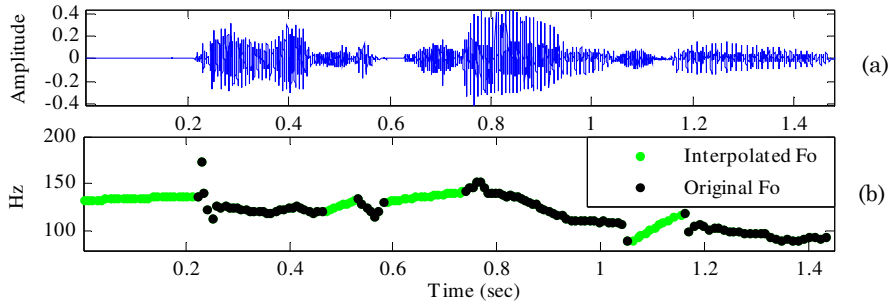


Figure 5.23: (a) A speech utterance (*Fs=16* kHz) and (b) original *F0* contour (black) and linearly interpolated *F0* contour (green).

5.4.3.2 Phrase and Accent Command Extraction

The processed F_0 contour is then used to estimate the phrase and accent events. The detection of phrase events or phrase boundaries is based upon the work reported in in [**205**]. The F_0 contour is lowpass filtered and the negative-to-positive transition of the derivative of the filtered F_0 contour is taken as phrase boundaries. The *strength* of the phrase boundary was estimated by the *slope* of the line at the negative-to-positive crossings. Accent commands parameter extraction is based on the work carried out in [**206**]. The procedure is to detect the largest maximum and the smallest minimum for each interval where the sign of the derivative of F_0 remains same. A pair of maximum and minimum corresponds to the *onset* and the *offset* of an accent command. An example of the Fujisaki parameters extracted from a speech utterance at *16* kHz is shown in Figure 5.24. It is observed that model generated contour approximates smooth version of original F_0 contour.

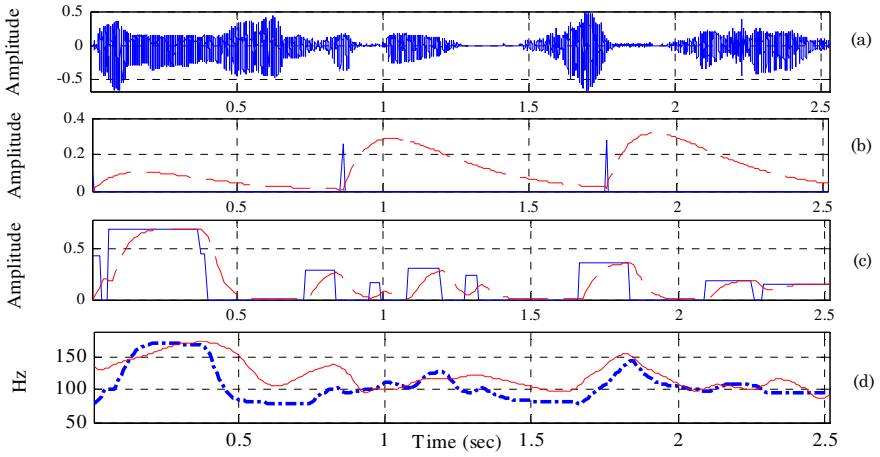


Figure 5.24: (a) Speech signal, (b) phrase commands (blue) and phrase components (dashed), (c) accent commands (blue) and accent components (dashed) and (d) original F_0 contour and model generated F_0 contour (dashed). Adapted from [207**].**

5.4.4 Estimation of the Vocal Fold Length

In this Section, we present our work on estimating the vocal fold length from the Fujisaki model and the *SoE*. This work is presented as a possible application of the Fujisaki model and is not applied for the present problem of spoof detection. The length of the vocal folds for male, female is found to be in the range of *17-25 mm*, *12.5-17.5* mm, respectively [**178**], [**208**]. For infants, this range is approximately *6-8 mm* and at birth the membranous length of the vocal folds (actually vibrating part)

is around *2 mm*. The location of the vocal folds makes the length estimation task even more challenging. Estimating the actual vocal fold length from the speech signal is difficult as the length of the folds changes while opening and closing.

5.4.4.1 Relation between L, F_0 and Stress

If we assume the vocal folds as 'ideal strings' with uniform properties, then the corresponding F_0 as per eq. (5.27) is given by [**178**],

$$Fo = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}}, \tag{5.34}$$

where L is the length of the vocal folds, σ is the longitudinal stress and ρ is the tissue density. Thus, knowing F_0, σ and ρ could give an *estimate* of the length of the vocal folds. There are two aspects to be considered here. First, F_0 is not constant along an utterance rather it varies. It is known from Fujisaki model that the F_0 contour is the result of the movement of *intrinsic* muscles in the larynx [**202**]- [**203**]. This gives a representation of the F_0 contour in log-domain as the *superposition* of two mutually independent contributions that occurs due to the independent movement of the thyroid cartilage and muscular reaction times [**201**]- [**202**]. Thus, the F_0 contour is the *superposition* of phrase components and accent components with a constant value F_b (minimum value of the speaker's F_0). Here, we use the model generated F_0 contour by Fujisaki model to get the F_0 contour. Second, the stress (σ) at the vocal folds also varies depending on various factors like presence of vowels or consonants, behavior/emotions of the speaker (e.g., angry and shouted speech will result in more stress on the vocal folds), prominence (focus or stress on one particular word affects neighboring words), etc. [**209**]. Therefore, to estimate stress (i.e., the force with which vocal folds close suddenly during the return phase), we estimate the strength with which the vocal folds close during the return phase of each glottal cycle. The locations of the *sudden* closure of vocal folds are the location of the impulses that excite the vocal tract as a system. However, the force with which the fold closes generates impulses with *varying* amplitude to excite the system. We use the ZF method to obtain the GCIs [**182**] and *SoE* is obtained by estimating the slope of the ZF filtered signal at GCI instants [**179**]. Thus, by the model generated F_0 contour and the *SoE* for an utterance, we obtain the running estimate of the vocal fold length.

5.4.4.2 Mathematical Derivation

Taking log (.) on both sides of eq. (5.34), we get,

$$\log F_0 = \log\left(\frac{1}{2L}\sqrt{\frac{\sigma}{\rho}}\right) \Rightarrow \log F_0 = \log\left(\frac{1}{2L}\right) + \frac{1}{2}\log\left(\frac{\sigma}{\rho}\right). \quad \textbf{(5.35)}$$

Simplifying eq. (5.28) of Fujisaki model and taking $T_0=o_o$, we get,

$$\log F_0 = \frac{1}{2}\log(c_0{}^2\,\frac{\sigma_0}{\rho}) + \frac{1}{2}bx. \quad \textbf{(5.36)}$$

Equating the eq. (5.35) with eq. (5.36) of log (F_0), we get,

$$\log\left(\frac{1}{2L}\right) + \frac{1}{2}\log\left(\frac{\sigma}{\rho}\right) = \frac{1}{2}\log(c_0{}^2\,\frac{\sigma_0}{\rho}) + \frac{1}{2}bx, \quad \textbf{(5.37)}$$

$$\therefore \quad \frac{1}{2}\log\left(\frac{\sigma}{c_0{}^2\sigma_0}\right) + \log\left(\frac{1}{2L}\right) = \frac{1}{2}bx, \quad \Rightarrow \log\left(\frac{1}{c_0}\frac{1}{2L}\sqrt{\frac{\sigma}{\sigma_0}}\right) = \frac{1}{2}bx, \quad \textbf{(5.38)}$$

$$\therefore \quad \frac{1}{c_0}\frac{1}{2L}\sqrt{\frac{\sigma}{\sigma_0}} = e^{\frac{1}{2}bx}. \quad \textbf{(5.39)}$$

Now, by the Fujisaki Model, the *constant term* F_b is

$$F_b = c_0\sqrt{\frac{\sigma_0}{\rho}}, \;\Rightarrow\; c_0\sqrt{\sigma_0} = F_b\sqrt{\rho}. \quad \textbf{(5.40)}$$

Using $c_0\sqrt{\sigma_0} = F_b\sqrt{\rho}$ in eq. (5.39), we get, $\frac{1}{2LF_b}\sqrt{\frac{\sigma}{\rho}} = e^{\frac{1}{2}bx}$

$$\therefore \; L = \frac{1}{2e^{\frac{1}{2}bx}F_b}\sqrt{\frac{\sigma}{\rho}}. \quad \textbf{(5.41)}$$

This gives the equation of length (L) of the vocal folds in eq. (5.41). The length L is directly related to stress along the folds. The term $bx/2$ is a summation of phrase and accent components (from eq. (5.28) and eq. (5.32)), F_b is constant for an utterance and ρ is assumed constant with a value *1* gm/cm^3. Figure 5.25 shows block diagram representation of the proposed method for estimating the length of the vocal fold from the speech signal.

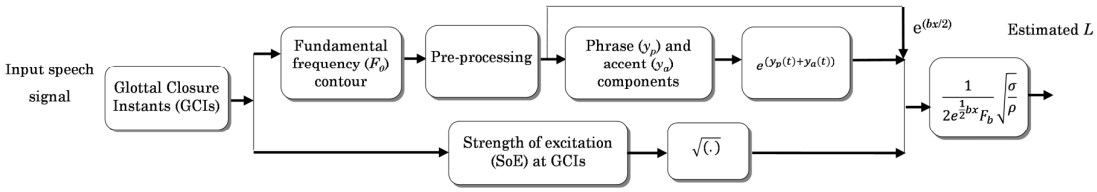


Figure 5.25: The block diagram of the proposed method for vocal fold length (*L*) estimation. Adapted from [210**].**

The procedure for estimating L is as follows:

- Estimate GCI locations and obtain the F_0 contour,
- Pre-process the F_0 contour as in Section 5.4.3.1 and estimate x_p and x_a,
- Pass the *phrase* commands (x_p) and *accent* commands (x_a) through h_p and g_a, to get phrase components (y_p) and accent components (y_a), respectively.
- Estimate the following:
  - F_b (minimum value of F_0 contour),
  - Obtain the SoE and find $\sigma = \sqrt{SOE}$,
  - $\exp(bx/2) = \exp((y_p(t) + y_a(t)))$.
- Finally, using $\rho=1$ gm/cm^3 in eq. (5.41) gives a running estimate of the vocal fold length.

The SI units of parameters in the eq. (5.41) are as follows,

$$Stress = Force / Area = mass \times acceleration / Area = kg \times ms^{-2} / m^2, \quad \textbf{(5.42)}$$

Density=mass/volume=kg/m^3, F_b=Hz, b has unit length^{-1} and x is change in length. Thus, exp($bx/2$) is constant. Therefore,

$$L(unit) = \frac{\left(kg.m.s^{-2}/m^2\right)^{1/2}}{Hz.(kg.m^{-3})^{1/2}} = \frac{kg^{1/2}.m^{-1/2}.s^{-1}}{Hz.kg^{1/2}.m^{-3/2}} = \frac{s^{-1}}{Hz.m^{-1}} = m\ (metre). \quad \textbf{(5.43)}$$

To study the length estimated from eq. (5.41), we initially take an utterance from the CMU-ARCTIC database [**211**] for male and female speakers. There are two ways in which the length is estimated. First, by directly using eq. (5.41) (L1) and second assuming that length changes only during voicing and remain constant during unvoicing, i.e., we assume length at unvoiced regions as an average value of length during voiced regions, i.e., replace length at the unvoiced regions by mean of length at the voiced regions (L2). Hence, for unvoiced sounds, the length of vocal folds is assumed to be constant. Figure 5.26-5.27, shows an example of estimation of L on

same speech utterance for male and female speaker, respectively. The red bars in the Figure 5.26(a)-5.27(a) show voiced regions. Figure 5.26(b)-5.27(b) is the F_0 contour obtained by Fujisaki model and Figure 5.26(c)-5.27(c) is the *SoE* of speech utterance in Figure 5.26(a)-5.27(a), respectively.

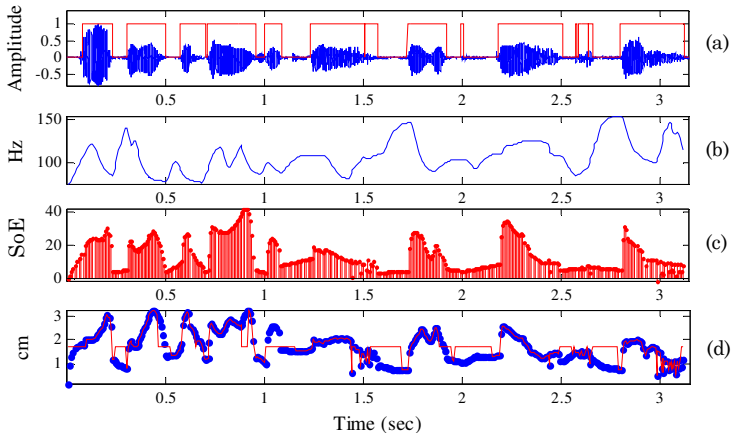


Figure 5.26: (a) Speech signal for a male speaker, (b) model generated F_0 contour, (c) *SoE* and (d) estimated vocal fold length (blue dotted) and length by replacing unvoiced regions by mean of length in voiced regions (red continuous). Adapted from [210].

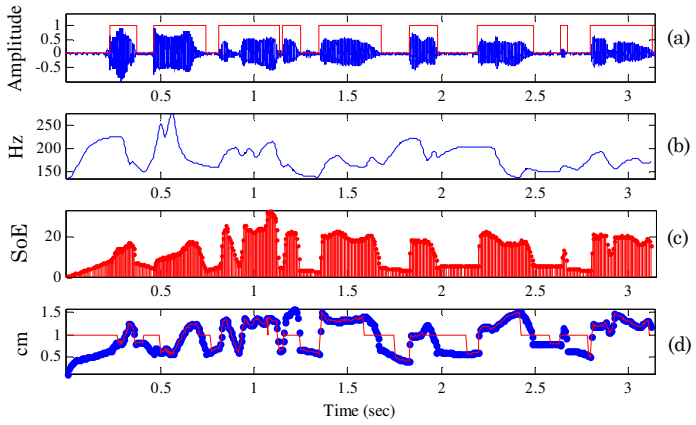


Figure 5.27: (a) Speech signal for a female speaker, (b) model generated F_0 contour, (c) *SoE* and (d) estimated vocal fold length (blue dotted) and length by replacing unvoiced regions by mean of length in voiced regions (red continuous). Adapted from [210].

In Figure 5.26(d)-5.27(d), the dotted line is an estimate of the vocal fold length by using eq. (5.41), i.e., L1, while the red continuous line is the length obtained after replacing the unvoiced regions by the mean of length in the voiced regions (L2). A similar analysis is carried out to estimate the vocal fold length of an infant as well. Figure 5.28 shows the length estimated for cry of an infant at birth. For cry, the unvoiced regions hardly occur as cry is mostly due to vibration of the vocal folds. The

average length for male, female and infants after replacing the unvoiced regions by mean of length of voiced regions (L2) was *18 mm*, *10 mm* and *2 mm,* respectively.

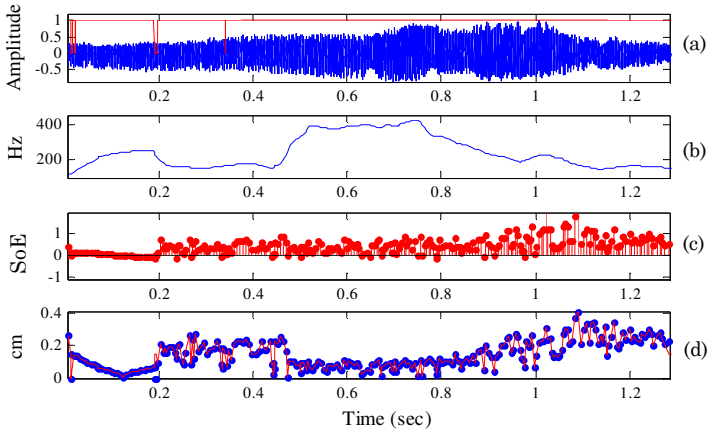


Figure 5.28: (a) Infant cry, (b) model generated *F0* contour, (c) *SoE* and (d) estimated vocal fold length (blue dotted) and length by replacing unvoiced regions by mean of length in voiced regions (red continuous).

Few experiments were carried out on all speakers (*2* female and *5* male) from the CMU-ARCTIC database [**211**]. As shown in Table 5.14, there is a difference of about *1-4 mm* between the estimated average lengths from L1 and L2 method. In the case of female speech, the average estimated length falls in the required range. However, for *rms*, *ksp* and *jmk* male speakers, the estimated length was high. It was observed that the *SoE* in the case of these speakers was high. Therefore, it might be possible that the present approach of estimating *SoE* might not be feasible for all speakers. In addition, all the utterances were sampled at *8 kHz* as *SoE* was very high for higher sampling frequency.

Table 5.14: The vocal fold length estimated for *2* female and 5 male speakers of CMU-ARCTIC database. Adapted from [210**]**

Method	Length of folds in *mm*						
	Female			Male			
	slt	clb	bdl	Rms	ksp	awb	Jmk
L1	15.5	20.6	16.1	35.1	27.8	20.1	56.3
L2	14.3	19.1	15.2	32.9	26.0	18.5	52.1

Thus, the proposed method captures variations in vocal fold length along an utterance and also the mean value can be a good estimate of the actual vocal fold length. However, there was no ground truth available and not all estimated lengths were in range. In this Section, we presented an approach to estimate the length of the vocal folds. However, the estimated length is at times larger than the intended

range and hence, the vocal fold measure is not used further for natural *vs*. spoofed speech analysis. Instead, we use the parameters derived from the Fujisaki model to analyze the natural *vs*. spoofed speech.

5.4.5 Fujisaki Model Parameters for Analysis of Spoofed Speech

Using the Fujisaki model parameters directly as features is not possible due to the varying lengths of the commands and components. In this Section, we discuss our initial work using the Fujisaki model parameters for analysis of natural and synthetic speech in the Gujarati language [**207**]. Figure 5.29 shows the spectrogram for natural speech, USS and HTS-based synthesized speech for the same utterance. The spectrogram of USS-based synthesized speech is similar to the natural speech in terms of speaker characteristics. However, there are breaks in the spectrogram representing discontinuity in the formant contour (dotted oval showing abruptness due to concatenation). These breaks may also occur in natural speech however, the frequency of their occurrences in USS-based speech is relatively more due to concatenation of speech sound units. The spectrogram of HTS-based speech shows loss in intelligibility and the formant structure do not appear to be preserved in HTS-based speech (dotted squares).

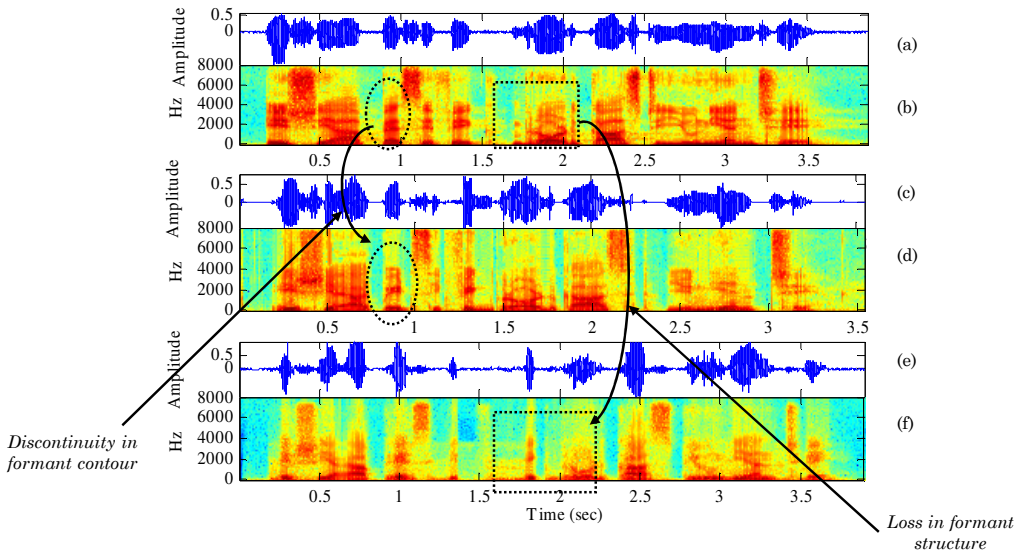


Figure 5.29 (a) Speech Signal, (b) spectrogram of (a), (c) USS-based speech, (d) spectrogram of (c), (e) HTS-based speech and (f) spectrogram of (e).

We quantify the difference in various spectrograms by the *Itakura–Saito* distance measure. It measures *perceptual* difference between an original spectrum $P(\omega)$ and its

approximation $\hat{P}(\omega)$. We consider synthetic speech to be an approximation to the natural speech. The utterances were time-aligned using DTW and the LPCs were extracted from the speech signal for every *20* ms speech frame with a frame shift of *10* ms for computation of IS distance, given by [**212**],

$$D_{IS}(P(\omega),\hat{P}(\omega)) = \frac{1}{N}\sum_{m=1}^{N}\left[\frac{P(\omega_m)}{\hat{P}(\omega_m)} - \log\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) - 1\right], \quad \textbf{(5.44)}$$

where N is the number of speech frames. Table 5.15 shows that the IS distance is relatively less between natural and USS speech as compared to HTS speech because IS distance measures spectral characteristics reliant on the *size* and *shape* of the vocal tract of the individual.

Table 5.15: Average IS distance between natural and synthetic speech over *100* utterances for male speaker and female speaker.

D_{IS}	USS		HTS	
	Male	**Female**	**Male**	**Female**
IS	11.675754	9.2565341	14.994685	18.448603

Initially, the analysis of the Fujisaki Model parameters was carried out on a small set of *100* utterances of natural, USS and HTS system for a male speaker and female speaker. Same text material is used for both natural and synthetic speech. The analysis was done using the parameters generated from the model generated F_0 contour in log-domain, i.e., F_b, x_p, y_p, x_a and y_a. The results of the analysis in terms of mean and standard deviation is presented in Table 5.15 and detailed descriptions are given in the following Sections. The present analysis is done on the utterances using the same text and later we extend the analysis to classification task on a generalized non-parallel and statistically meaningful dataset.

5.4.5.1 Minimum Value of F_0 Contour (F_b)

In the representation of Fujisaki model, F_b is the baseline value of F_0 contour. It is a constant term $c_o(T_o/o)^{1/2}$ as long as the speaker maintains *same* speaking style and emotional state [**203**]. The F_b value for female speaker is higher than male speaker for natural, USS and HTS voices (as in Table 5.16). For USS, the speech sound units are concatenated using units from the same speaker. Therefore, the mean value of F_b is nearly same to the natural speech. However, as the units are concatenated from several sessions of recording, there exists more variability in the F_b (more standard

deviation (*sd*)). More the professional artists are consistent in recording, lesser would be the variation in F_b. For HTS, the mean F_b will be close to the natural speech if naturalness in HTS speech is preserved. Other direct inferences for *sd* could not be drawn and hence, phrase and accents components are used for further analysis.

Table 5.16: The distribution (in terms of the mean and standard deviation (*sd*) for *100* utterances (natural, USS and HTS) for a male and female speaker) of the minimum value of F_0 contour (F_b), phrase components (y_p) and accent components (y_a). Adapted from [207].

Parameters of MModel	Natural				USS				HTS			
	Male		Female		Male		Female		Male		Female	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
F_b	64.79	6.045	123.1	19.37	67.7	8.788	131.9	22.31	65.51	10.02	92.69	14.13
y_p	0.115	0.105	0.389	0.272	0.148	0.108	0.314	0.229	0.179	0.128	0.245	0.196
y_a	0.339	0.250	0.375	0.281	0.279	0.199	0.340	0.231	0.290	0.165	0.426	0.312

5.4.5.2 Phrase Commands and Phrase Components

The instant when the *cricoid* muscles undergo a translation motion, an *impulse* is generated corresponding to *phrase breaks* which occur naturally during speech production. Such prosodic breaks are not more prominent in synthetic speech. Figure 5.30 shows the number of breaks in the natural and the synthetic speeches. In USS, silences were present in synthetic speech as per the text punctuations, whereas during natural speech production, prosody is automatically generated as per the nature of the utterance, context, etc. Therefore, USS synthesized speech is expected to have less or equivalent number of *phrase* breaks as compared to the natural speech as seen in Figure 5.30. For HTS male and female, the number of phrase breaks increases. Next, impulses due to phrase commands are passed through 2^{nd} order phrase control mechanism, to produce the phrase components. The USS synthesized speeches have similar means and *sd* for phrase components to that of

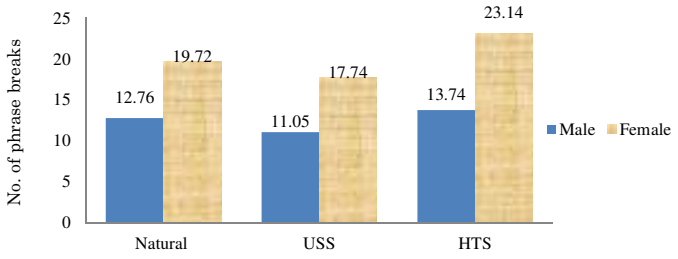


Figure 5.30: Number of phrase breaks in natural, USS and HTS speech. Adapted from [207].

natural utterances (as in Table 5.16). In HTS, for the number of phrase breaks was more. However, their strength was less due to the fact that these breaks were due to

phrase pauses decided from the text and not due to an actual change in translational motion of the cricoid muscles. Hence, the mean value of phrase component is less.

5.4.5.3 Accent Commands and Accent Components

The accent parameters of the Fujisaki model capture the variations in the speech which includes the stress that is applied to a particular word, syllable, etc. Especially for interrogative and exclamatory types of sentences, the accent parameters vary significantly. In the case of USS speech, due to concatenation, the natural variation due to stress on a syllable, word etc. may not always be present. Hence, the synthetic speech sounds monotonous while listening and this brings a possibility for the accent commands and components to vary less than natural speech signal. Thus, for USS synthesized speech, the mean and *sd* of the accent components was less than the natural speeches (as shown in Table 5.16). In the case of HTS, any uniform pattern for the accent parameters was not found for either male or female.

5.4.5.4 Statistical Analysis of Results

The scatter plots for *100* USS and *100* HTS synthesized utterances formed by the mean of accent and phrase components are shown in Figure 5.31. The clusters for USS and natural speech are different in size and shape than HTS and natural speech. In particular, clusters for USS synthesized and natural speech are found to be more overlapping and it is difficult to identify a boundary for these two classes. On the other hand, for HTS speech (especially, female voice), the two classes are relatively better *separable*. Thus, the female voice in HTS *lacks* relatively the prosodic features as that of the natural voice.

To know the difference in the distribution of the parameters for the natural and synthetic voices, we performed the Student's *t-test* to investigate if the two sets of synthetic voices are significantly different from natural. It is seen from Table 5.17 that the *null* hypothesis for all the synthetic voices is *rejected* with very less probability (<than *0.0001*) in most cases. Hence, the natural and synthetic systems (USS and HTS) have diverse means, which is effective while training statistical models like GMM, etc. This shows that the phrase and accent parameters could prove a good set of features to distinguish natural and synthetic speech.

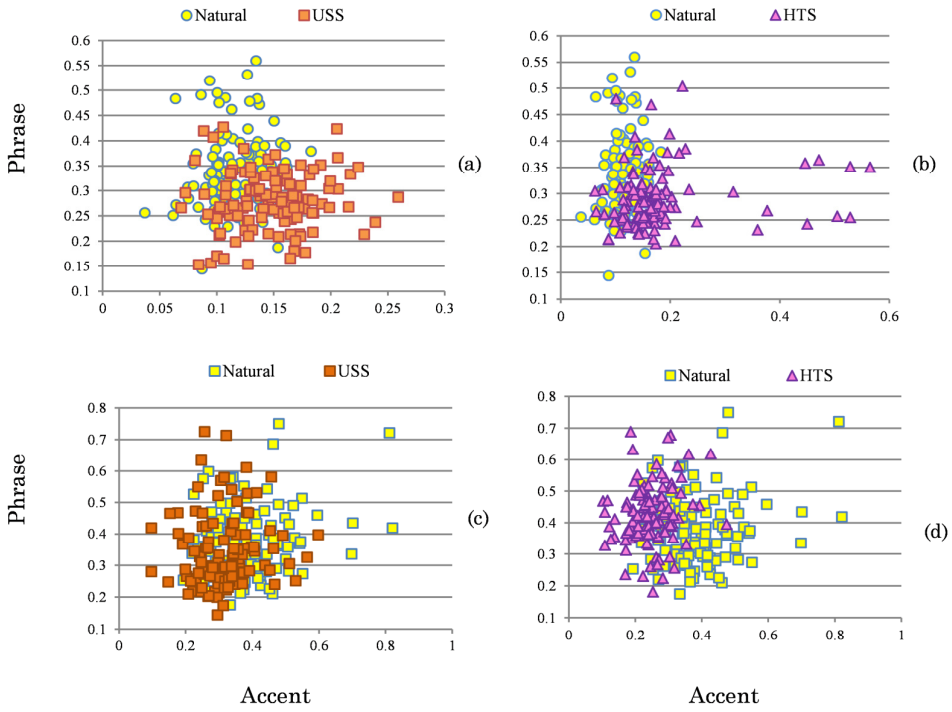


Figure 5.31: Clusters of accent and phrase components for (a) natural *vs.* USS (male), (b) natural *vs.* HTS (male), (c) natural *vs.* USS (female) and (d) natural *vs.* HTS (female). Adapted from [207].

Table 5.17: Probability of rejecting the null hypothesis for phrase and accent parameters in USS and HTS. Adapted from [207].

System	USS		HTS	
	Phrase	Accent	Phrase	Accent
Male	<0.0001	<0.0001	<0.0001	<0.0001
Female	<0.0001	0.018	<0.0001	0.0002

For both USS and HTS, it was observed that phrase command could serve as an important feature to distinguish between natural and synthetic speech. For HTS speech, phrase components could be effective. Results of *t-test* show that the null hypothesis is rejected in all the cases, which is effective while training statistical models for classification problem on large dataset as discussed next.

5.4.6 Experimental Results

The design of features from the Fujisaki model parameters is a difficult task as the information of the prosodic phrase breaks and accent components lies in the location of their occurrence and is usually observed when the utterances are parallel. Therefore, it is generally not so easy and rather challenging to generalize these

features for the non-parallel utterances. In this Section, we develop a feature vector comprising of the F_0 and the estimated phrase and accent parameters. This feature set is then evaluated on the ASV spoof challenge database and the Blizzard datasets.

5.4.6.1 Parameterization

In the estimation of the phrase and accent parameters, first, the F_0 is estimated from the ZF filtering method using a frame size of *25* ms and a *50* % overlap. The phrase and accent components are estimated as discussed in Section 5.4.3.2. In addition to the F_0 estimated from the ZF method, we consider the model generated F_0 contour generated from the phrase and the accent components as well. The extraction of features from the phrase and accent commands is difficult as these are available at only particular instant or for a fixed duration of time, respectively. These time instants are different for different speech utterances. Therefore, we use the phrase and accent components as features. To form a feature vector, we take the F_0 generated by ZF method, the model generated F_0 (MF_0), the phrase components (y_p) and the accent components (y_a). In addition, we obtain the dynamic variations across these time-varying representations to form an *8-D* feature vector comprising of F_0, ΔF_0, MF_0, ΔMF_0, y_p, Δy_p, y_a and Δy_a for the SSD task.

5.4.6.2 Results on the Development Set of ASVspoof challenge Database

Due to the small feature dimension, we consider an initial experiment to estimate the % EER of Fujisaki model-based features for various Gaussian mixture components. The mixture components are varied from 2, *4*, *8*, *16*, *32*, *64* and *128*. It is observed that the general trend is a decrease in EER with the increase in the mixture components. As observed in Figure 5.32, the least EER of *41.19* % is obtained for *128* mixture components. Hence, as in all the previous experiments to maintain uniformity as well, we use *128* mixture components in this work.

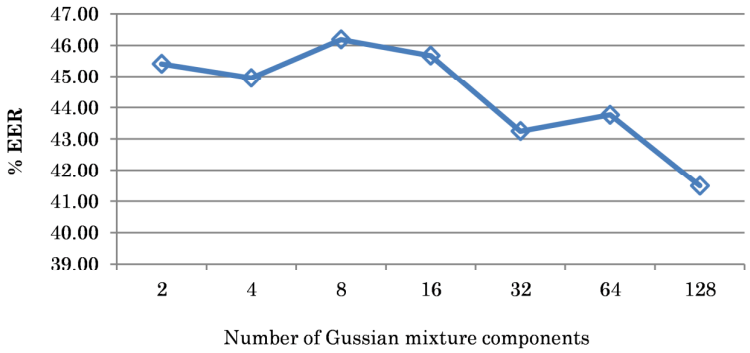


Figure 5.32: The % EER obtained on the development set for the Fujisaki model-based features at varying number of Gaussian mixture components.

Next, we obtain the performance of the Fujisaki model-based features on the development set of the ASV spoof challenge database. As observed in Table 5.18, the average % EER obtained using *128* mixture components is very high around *41.493 %*. It should be noted that that the development set consists of only vocoder-based spoofs and for such a case, this % EER obtained is very high. We also attempt to perform a score-level fusion of the Fujisaki model features and the system-based MFCC, CFCC, CFCCIFS and SBAE features to find out possible complementary information. However, as shown in Table 5.18, even for score-level fusion with any weight factor, the system-based features do not show any significant improvement as that obtained with the F_0, *SoE1*, *SoE2* and prediction-based features. Hence, this proposed feature vector needs to be modified in an efficient manner as to make them useful for the SSD task.

Table 5.18: EER (in %) for score-level fusion of Fujisaki model-based feature set with the system-based feature sets (using *D3* feature vector) at various fusion factors a_f on the development set

Feature Set 1	Fusion Factor (a_f)											Feature Set 2
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
Fujisaki Model Features	41.493	15.613	7.807	4.461	3.117	2.373	1.945	1.716	1.630	1.630	1.601	**MFCC**
	41.493	14.984	7.292	4.118	2.717	2.116	1.687	1.544	1.487	1.516	1.544	**CFCC**
	41.493	20.103	8.693	4.432	2.717	1.859	1.487	1.344	1.258	1.201	1.230	**CFCCIFS**
	41.493	17.215	9.580	5.633	3.632	2.602	2.002	1.659	1.544	1.487	1.487	**SBAE**

Score-level fusion is carried as per eq. (3.6)

Dependency on spoofing algorithms: Figure 5.33(a) shows the spoof dependency of the Fujisaki model-based features on the development set. It is observed that the trend is similar to that observed in F_0, *SoE1* and *SoE2* features, i.e., except S2, the other spoofs when used alone in training gave less % EER for known and the same

type of attacks. However, for the known and same type of attacks, the EER is > *15* % which is very high as compared to that obtained by other source-based features. The lack of the feature set to model the spoof-specific characteristics is evident from the high % EER. For different type of spoofs, the detection is very poor resulting in very high ~*70* % EER. The *S2* spoof showed a rather opposite behavior which needs to be investigated further.

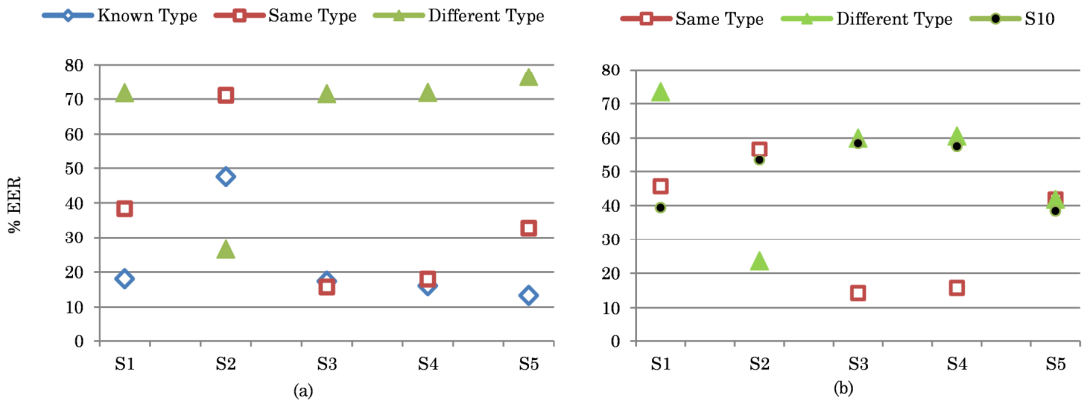


Figure 5.33: The % EER on known, same type, different type and *S10* attack when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for Fujisaki model-based feature set and tested on the (a) development set and (b) evaluation dataset.

5.4.6.3 Results on the Evaluation Set of ASVspoof challenge Database

The results on the evaluation set for the features estimated from the Fujisaki model is shown in Table 5.19. The EER with *8-D* Fujisaki model-based feature set was *43.11* %. The individual % EER for *S1* spoof was the least with *27.42* % while for *S1-S9* the EER is between *40-50 %*. It is observed that even after the score-level fusion with the system-based features, the % EER does not decrease and hence, in the current form, the feature set do not capture significant spoof-specific characteristics.

Table 5.19: EER (in %) for score-level fusion of Fujisaki model-based feature set with the system-based feature sets (using *D3* feature vector) at various fusion factors a_f on the evaluation set

Feature Set 1	Fusion Factor (a_f)											Feature Set 2
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	
Fujisaki Model Features	43.11	18.89	10.05	6.68	5.22	4.66	4.41	4.26	4.25	4.23	4.264	**MFCC**
	43.11	17.45	7.72	4.15	2.80	2.20	1.88	1.77	1.74	1.74	1.755	**CFCC**
	43.11	22.22	10.30	5.17	3.11	2.19	1.79	1.66	1.62	1.61	1.606	**CFCCIFS**
	43.11	22.18	12.04	7.06	4.71	3.40	2.83	2.58	2.47	2.46	2.488	**SBAE**

Dependency on spoofing algorithms: Figure 5.33(b) shows the spoof-dependency of the Fujisaki model-based features on the evaluation set. It is observed that SS spoof could identify its same type with around *15* % EER, its different type, i.e., VCS and *S10* spoof with ~ *60* % EER. On the other hand, for the training with VCS spoof, no uniform conclusion could be drawn. Hence, the features using the prosodic information need to be further investigated to derive significant conclusions.

5.4.6.4 Results on the Blizzard Challenge 2012 Database

The results of the performance of the Fujisaki model-based features on the Blizzard Challenge 2012 and Blizzard Challenge 2014 database are shown in Table 5.20. It is observed that unlike the other source-based features, the prosodic information in terms of phrase and accent did not show any significant differences even for the unit-selection or HMM-based speech synthesis systems. For Blizzard 2012 dataset, the % EER for USS-based speech and HMM-based speech was ~*40-50* %. The hybrid and diphone systems showed more EER of around ~*60* %. For Blizzard 2014 dataset, the % EER for all Gujarati systems was > *60* %. Similar observations were obtained for the Hindi language as well.

Table 5.20: EER (in %) for Fujisaki model-based features sets on training with the ASV spoof data and testing with the Blizzard Challenge databases

Blizzard 2012	English Systems	% EER	Blizzard 2014	Gujarati systems	% EER	Blizzard 2014	Hindi systems	% EER
USS	**B**	43	**HMM**	**C**	70	**HMM**	**B***	65
Hybrid	**C**	48	**HMM**	**D**	69	**HMM**	**C**	53
Hybrid	**D***	60	**HMM**	**E**	59	**Hybrid**	**D**	62
HMM	**E***	39	**HMM-DNN**	**F**	79	**HMM**	**E**	51
USS	**F**	54	**USS**	**G**	62	**HMM-DNN**	**F**	68
USS	**G**	41	**HMM**	**H**	76	**USS**	**G**	36
HMM	**H**	45				**HMM**	**H***	67
USS	**I**	49				**HMM**	**K**	56
Diphone	**J***	64						
HMM	**K***	50						

* systems with lower MOS from $1 \leq 2$ (wavefiles for Gujarati system B and system I and Hindi system I are not available)

5.4.7 Summary

In this Section, we attempt to design a low-dimensional feature vector from the parameters estimated from the Fujisaki model. In the literature, the prosodic difference in terms of phrase and accent parameters was studied in the context of

parallel utterances. Here, we assume that the spoofed speech lacks the prosodic characteristics as that are present in the natural speech and try to generalize the findings from the phrase and accent components to non-parallel utterances. Firstly, this does not apply always for all utterances as even natural speech can be spoken without significant prosodic variations. Secondly, if a prosodic model has been applied while speech synthesis, then in such a case it is difficult for these features to detect spoof-specific characteristics. Hence, there is a need to modify and develop features that perform significantly well on the ASV spoof challenge database and then study its generalization.

5.5 Chapter Summary

In this Chapter, we discuss several approaches to distinguish natural *vs.* spoofed speech using source-based features. The source modeling is generally not as much explored as the system-based feature. Hence, they gave complementary information as compared to the system-based features. The system-based features worked well for vocoder-independent spoof and the source-based features perform better for vocoder-based speech, hence, the overall performance of the detection system improves with their score-level fusion. In the next Chapter, rather than using source and system separately, we explore the features that have embedded both information about the excitation source and vocal tract system (i.e., filter). With respect to this, we explore the nonlinear source-filter interaction or coupling between the two to capture the spoof-specific characteristics for the SSD task.

Chapter 6.
Source-Filter Interaction Features

6.1 Introduction

In this Chapter, we discuss a crucial aspect of the speech production mechanism, i.e., the nonlinear Source-Filter (S-F) interaction. It is known that there exists nonlinear interaction between the excitation source and the vocal tract as the system. This interaction is an attribute of natural speech and is not present in Synthetic Speech (SS) or Voice Converted Speech (VCS). This Chapter proposes features based on S-F interaction for the Spoofed Speech Detection (SSD) task. To that effect, we estimate the voice excitation source (i.e., the glottal flow derivative waveform, $\dot{g}(t)$) and model it using well-known Liljencrants-Fant (LF) model to get the coarse structure $g_c(t)$. The residue or the difference, $g_r(t)$, between estimated $\dot{g}(t)$ and fitted model $g_c(t)$ captures this nonlinear S-F interaction. Two approaches have been proposed, i.e., the time-domain approach and frequency-domain approach. The features are evaluated on the ASVspoof 2015 challenge database and we explore the features for robustness in the presence of additive white noise at various Signal-to-Noise Ratio (SNR) levels and for channel mismatch conditions on the Blizzard Challenge datasets.

6.2 Basis for the Proposed Approach

The vocal folds along with their dynamic movement represent various aspects of both speech and the speaker. During phonation, the gradual opening of the glottis and its sudden closure results in an *asymmetric* shape of the glottal flow waveform ($g(t)$). Assuming a Linear Time-Invariant (LTI) speech production mechanism, the derivative (due to lip radiation [**21**], [**177**]) of the glottal flow waveform $g(t)$ is referred to as the voice excitation source (i.e., $\dot{g}(t)$). This voice excitation source can be parameterized using physical or acoustic models. Physical models such as the two-mass model of Ishizaka and Flanagan involve the use of a large number of independent parameters for modeling [**213**]. The acoustic LF model is also a good approximation to $\dot{g}(t)$ and it can be represented in the frequency-domain as well

[**214**]. The LF-model gives the shape and timing parameters of the voice excitation source which are known to relate to voice quality measures such as speed quotient (SQ), open quotient (OQ) and return quotient (RQ) [**215**]. Natural speech has variations ranging from creaky to breathy voice which needs to be incorporated in synthesis and voice conversion techniques for better voice quality. In this context, incorporating the excitation source information through LF-model into HMM-based synthesizer has shown to give more naturalness and reproduction of two basic voice qualities such as breathy and tense [**216**]. Another example is of GlotHMM, which uses inverse filtering for generating glottal excitation and modeling it in an HMM framework using Line Spectral Frequencies (LSFs) [**184**]. Other earlier source models in parametric speech synthesizers include simple pulse and noise excitation model [**33**], Multi-Band mixed Excitation (MBE) [**217**], STRAIGHT vocoder that uses a mixed excitation model [**218**], Harmonic plus Noise Model (HNM) of speech [**219**], etc. The use of system-level features to model the vocal tract filter is very well known in speech synthesis and voice conversion techniques. These system-level features include the MFCCs [**8**], generalized Mel-cepstral Coefficients (MCC) [**220**], [**221**], LSF representation of Linear Prediction Coefficients (LPC) [**222**], and approaches such as STRAIGHT-based speech parameters encoded into MCCs or LSFs. Thus, the majority of the existing techniques model the excitation source or system characteristics individually. However, from signal and systems respective, it is not only the independent role of the source or system features that contributes in producing natural speech, but it is also the time-varying dependencies between them or the nonlinear S-F *interaction* that effectively contributes to naturalness and speaker identity [**223**], [**224**].

6.2.1 The Source-Filter (S-F) interaction

In the linear S-F theory, the source of speech production is independent of the vocal tract (filter). In such a case, the source impedance is much higher than the input impedance to the vocal tract. However, due to the narrow constriction of the vocal tract above the glottis, there exists a nonlinear S-F interaction. The closer the constriction is to the vocal folds, greater is the degree of interaction [**225**]. In this case, the source impedance is comparable to the vocal tract input impedance. This makes the glottal flow highly dependent on the acoustic pressures in the vocal tract. According to the landmark investigations in [**225**], [**226**], [**227**], there are two

primary levels of S-F interaction, namely, Level 1 and Level 2. The Level 1 interaction occurs due to feedback from the vocal tract acoustic pressure (i.e., standing waves) that imparts variations in the glottal airflow (i.e., transglottal pressure drives the glottal flow which in turn is affected by the epiglottis pressure). On the other hand, Level 2 interaction primarily occurs in cases with high F_0 where the pitch harmonics are near the formants. It is responsible for variations in vocal fold vibrations (i.e., tissue movements) that occur with same pressure (i.e., intraglottal pressure drives the vocal folds) [**225**].

The findings in [**225**] summarize that there always exists an interaction of glottal airflow with the acoustic vocal tract. The primary effect of Level 1 interaction is the glottal airflow skewing (which in turn, balances source spectrum in terms of odd and even harmonics) that can be expressed by an analytic formula [**214**], [**184**]. The pressure from the vocal tract against the glottis will slow the flow and change its skewness. In addition to the asymmetric glottal flow, simulation of a simplified electrical first formant model showed the presence of a sinusoidal 'ripple' component (i.e., a fine structure superimposed onto the coarse structure) onto the open phase of the glottal flow [**21**]. Other effects of S-F interaction include abrupt increase in the first formant (F_1) and the corresponding *-3* dB bandwidth when the glottis opens. The increase in the bandwidth of the formant causes sudden decay of the vocal tract impulse response within a glottal cycle and is responsible for truncation effect in the speech waveform [**21**].

In [**228**], a structure for $\dot{g}(t)$ which includes both coarse structure and ripple component (produced as a result of S-F interaction) was proposed. The use of ripple (i.e., the fine structure features) is much explored in speaker identification task [**223**], [**229**]. However, its use in speech synthesis or voice conversion is not much explored yet. Thus, SS or VCS may not sound natural and intelligible than a certain extent due to the lack of nonlinear S-F interaction. To that effect, in this Chapter, we explore the fact that the nonlinear S-F interaction is an attribute of the natural speech production mechanism and not that of the machine-generated speech. It is highly complex for the synthetic speech generation or conversion systems to build or mimic such S-F interaction. With this motivation, we study the differences between the actual voice excitation source $\dot{g}(t)$ and it's coarse structure $g_c(t)$ (i.e., fitted LF-model). The voice excitation source $\dot{g}(t)$ is estimated by using a linear inverse

filtering technique. The use of a linear inverse filtering does not capture the effects of time variance in formant frequencies (due to the inherent segment-level block-based processing in the LP approach). That is, the true formants changes when the vocal folds are open *vs.* when they are closed are not captured. However, the anti-resonances of the glottal inverse filtering method remain fixed over many pitch (fundamental) periods. Thus, on using a linear fit to model the glottal flow, the anti-resonances that appear as time-varying amplitude variations in the open phase of the glottal flow are captured as ripple structures.

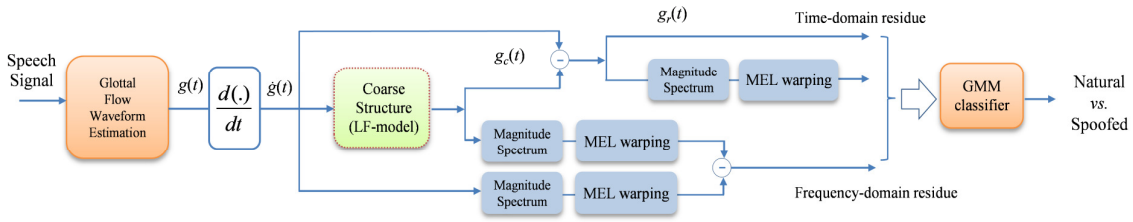


Figure 6.1: Schematic diagram of the proposed S-F interaction feature extraction process (both in time and frequency-domain) for the SSD task. Adapted from [93].

To fit the acoustic LF-model, an exhaustive search method is used to obtain $g_c(t)$ for an estimate of $\dot{g}(t)$ [**230**]. This approach varies the shape parameter R_d within a specific range and attains the best fit depending on the minimum cost obtained both in the time and frequency domains. This motivated us to consider the residuals both in time and frequency domains. The fitted LF-model $g_c(t)$, when subtracted from the $\dot{g}(t)$ gives the residual signal $g_r(t)$. The residual signal has information about the ripple (due to first formant (F_1) modulation of the vocal tract system) and the aspiration components (due to turbulence at the vocal folds) [**228**]. Thus, in the time domain, the L^2 norm of $g_r(t)$ in the closed phase, open phase and return phase of the glottis can be considered as feature representations for SSD task. In addition, as the first formant (F_1) modulation information is in the lower frequency range (*<1 kHz*), we also consider Mel representation (having high resolution for frequencies *<1 kHz*) of the residual $g_r(t)$. Furthermore, the residual information is also obtained in the frequency domain by using the difference between the spectrogram of $\dot{g}(t)$ and the spectrogram of $g_c(t)$. The residue in the frequency domain is further obtained by using the difference of the Mel wrapped spectrograms of $\dot{g}(t)$ and $g_c(t)$. Thus, shape and energy-based features in the time-domain and several feature representations in

the frequency-domain are used for the SSD task. The schematic of best representative features in time and frequency-domain are shown in Figure 6.1.

6.3 Voice Source Parameterization

The first task in this work involves estimating the excitation source $\dot{g}(t)$. To that effect, this section describes the inverse filtering approach to obtain $\dot{g}(t)$, followed by the description of the LF-model and its estimation from the R_d parameter using a search algorithm.

6.3.1 The Coarse Structure (LF-Model)

To obtain an initial estimate of voice excitation source, $\dot{g}(t)$, we assume the speech production mechanism as LTI system [**21**], i.e.,

$$s(t) \approx A\frac{d}{dt}\left[g(t) * h(t)\right] = A\left[\frac{d}{dt}g(t)\right] * h(t) = A\dot{g}(t) * h(t), \tag{6.1}$$

where * is the convolution operation, A is the gain that controls loudness, $s(t)$ is the speech signal, $\dot{g}(t)$ is derivative of the glottal flow waveform and $h(t)$ is the impulse response of the vocal tract system. Thus, to obtain an initial estimate of the $g(t)$, Iterative Adaptive Inverse Filtering (IAIF) method is used to inverse filter the vocal tract information from the speech signal [**183**]. From $g(t)$, its derivative $\dot{g}(t)$ is obtained to further characterize it in terms of the coarse structure $g_c(t)$. In the IAIF method, the effect of the vocal tract system and lip radiation is supressed from the speech signal to give an estimate of $g(t)$. The block diagram of IAIF method is shown in Figure 5.3 (as discussed in Chapter 5, Section 5.2.4).

The coarse structure $g_c(t)$ is a parameterization of the shape of the voice excitation source $\dot{g}(t)$. As shown in Figure 6.2, according to the time intervals, $g(t)$ is divided into *closed phase* (when the vocal folds are closed and ideally, there is no flow of air through the glottis), *open phase* (when the vocal folds are open) and *return phase* (when the glottis closes). The shape and flow of the regions represent different attributes of the speaker and the nature of speech signal. For example, in the case of shouted speech, the vocal folds close abruptly and hence, the return phase is much shorter [**21**]. The coarse component $g_c(t)$ of $\dot{g}(t)$ is modeled using the LF-model in the closed, open and return phase. Therefore, the LF-model shown in Figure 6.2 can be defined by the five time instants, namely, the glottal opening time (t_o), the instant

when the glottis closes (t_c), the time when $\dot{g}(t)$ crosses zero (t_p), the time when the $\dot{g}(t)$ reaches its maximum negative value (t_e) and the time when tangent to the return phase that crosses the time-axis (t_a). Similarly, we can define time periods corresponding to the LF-model as T_o (i.e., duration of a glottal cycle), T_p (period from t_o to t_p), T_e (period from t_o to t_e) and T_a (period from t_e to t_a). These attributes are primarily for voiced speech.

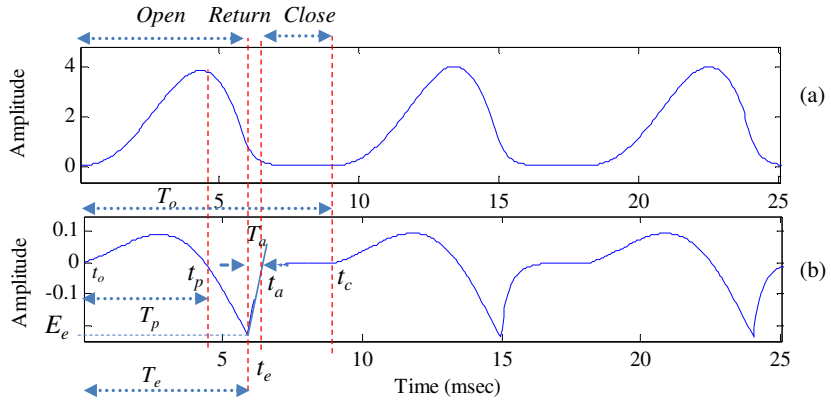


Figure 6.2: (a) Schematic of *g*(*t*) and (b) the corresponding derivative of the *g*(*t*) along with various timing instants and the time periods used in LF-model.

Thus, over a glottal cycle, the LF-model consists of an exponentially increasing sinewave, a decaying exponential function, and completed with a zero amplitude region, as described by the following equations [**214**], [**216**]:

$$g_c(t)|_{LF} = \begin{cases} e_{open}(t) = E_0 e^{\alpha t}\sin(\omega_o t), & t_o \le t \le t_e, \\ e_{return}(t) = E_1[e^{-\beta(t-t_e)} - e^{-\beta(t_c-t_e)}], & t_e < t \le t_c, \\ e_{close}(t) = 0, & t_c < t \le T_o, \end{cases} \tag{6.2}$$

where $E_0 = -E_e/(\sin(\omega_o t_e)e^{\alpha t_e})$ and $E_1 = -E_e/(1-e^{-\beta}(t_c - t_e))$ in eq. (6.2), parameters E_0 (amplitude of the sinewave), ω_o (angular frequency of the sinewave related to the rise time of the $g(t)$) and α (growth factor) determine the shape parameters in the open phase. The parameter E_e (amplitude of maximum excitation) and β (exponential time constant) constitute the shape parameters in the return phase. The time instant t_o is assumed to be zero and it is omitted in the formulas. The parameters of the LF-model can be obtained by considering the following assumption [**214**]:

$$\int_0^{T_0} g_c(t)dt = 0, \quad \therefore e_{open}(t_e) = e_{return}(t_e) = -E_e. \tag{6.3}$$

In a study reported in [**215**], a set of dimensionless parameters of the LF-model, called *R*-parameters, have been derived. These parameters can be expressed as dimensionless quotients often used to describe the shape of the glottal source signal, $\dot{g}(t)$. The *R*-parameters affect the coarse structure representation both in time and frequency domains. The *R*-parameters are given by,

$$R_g = \frac{T_o}{2T_p}\ ,\ R_k = \frac{t_e - t_p}{T_p},\ \textbf{and}\ \ R_a = \frac{T_a}{T_o}. \tag{6.4}$$

The parameters R_k and R_a are known to relate to the speed quotient (*SQ*) and return quotient (*RQ*), respectively. The *R*-parameters are related with the open quotient (*OQ*) as [**215**]:

$$OQ = \frac{1+R_k}{2R_g} + R_a. \tag{6.5}$$

In [**215**], an R_d parameter was developed that captured almost all possible variation of the LF-model. The R_d parameter is represented as,

$$Rd = 1000\left(\frac{E_0}{E_e}\right)\left(\frac{f_0}{110}\right), \tag{6.6}$$

where E_0 is the peak amplitude of the $g(t)$ and E_e is the maximum negative of $\dot{g}(t)$. The R_d parameter is related to R_g, R_k, and R_a by following approximation [**215**]:

$$R_d\ = \frac{1}{0.11}\left(0.5+1.2R_k\right)\left(\frac{R_k}{4R_g+R_a}\right). \tag{6.7}$$

Each of the *R*-parameters affects the coarse structure representation in the time-domain and its spectrum in the frequency-domain. There have been various approaches in the literature to determine the coarse structure of the estimated $\dot{g}(t)$. One of the approaches includes minimizing the least square error when $g_c(t)$ as in eq. (6.2) is fitted to the estimate of $\dot{g}(t)$. The error function in such a case is a nonlinear function of the model parameters and needs to be solved iteratively using nonlinear least square algorithm [**224**]. However, this approach depends on the initial estimates of the time and shape parameters. In addition, it is rather difficult to obtain accurately these timing parameters from the speech waveform. Thus, in this work, instead of using an iterative algorithm that optimizes all shape and time parameters of the LF-model, we consider the work carried out in [**230**], where only the R_d parameter (which is descriptive of all the shape parameters in the LF-model) is varied. This approach not only aims at minimizing the error in the time-domain

rather it minimizes the error in the frequency-domain as well. Moreover, we can estimate the other R-parameters from the R_d parameter [**215**].

6.3.2 Determination of GCI and F_0

In this work, we need to estimate the Glottal Closure Instants (GCIs) to fit the LF-model at each glottal cycle in the time-domain. Hence, we adapt the time-domain approach to estimate the GCI. The Zero Frequency (ZF) filtering method, also known as the *0-Hz* resonator is used [**182**] (as discussed in Chapter 5, Section 5.2.3). The basic idea behind ZF method is that the effect due to an impulse is spread uniformly across *all* the frequency regions including zero frequency. Thus, by passing the speech signal through the ZF filter, we try to decouple the interference of the vocal tract system (whose resonances are at much higher frequencies than zero frequency) from the excitation source. Therefore, to estimate the GCIs, the speech signal is passed through a ZF filter and the negative-to-positive zero-crossings of the filtered signal are hypothesized as an estimate of GCIs.

The procedure to fit the LF model to the glottal source, $\dot{g}(t)$, using an exhaustive search method and dynamic programming is done using the voice analysis toolkit [**231**] as given in the next sub-section. The accuracy of the GCI estimation algorithm will affect the F_0 estimate and the time period for which the LF-model is fitted to the glottal estimate. However, the search method for the R_d parameter and dynamic programming implementation uses estimated GCI locations from any given algorithm and then corrects the GCIs to match the main excitations of the $\dot{g}(t)$ (i.e., at its most negative peak in each glottal cycle). Thus, the GCI locations are aligned and adjusted to coincide with glottal source excitation minima. Hence, the dependency on the GCI extraction algorithm will be reduced. As an alternative to estimating the GCI locations from the speech signal, the $\dot{g}(t)$ estimated using the IAIF method can also be used directly for GCI estimation. However, this approach will require thresholding and peak picking of the negative peaks in the estimated $\dot{g}(t)$.

6.3.3 The Exhaustive R_d Search Algorithm

In [**230**], a search algorithm was proposed in which the LF-model is estimated for all possible R_d and the best R_d is searched that minimizes the error in time and

frequency domains. The steps to determine R_d are given in Algorithm 6.1 and its MATLAB implementation is available at [**232**]. In the estimate of frequency-domain error, H_g and H_c are the harmonic spectra of $\dot{g}(t)$ and $g_c(t)$, respectively.

Algorithm 6.1: Exhaustive search algorithm to estimate R_d. After [**230**].

Step 1	For each GCI centered frame $\Rightarrow R_d = 0.3 : 0.1 : 5$
Step 1a	Use F_0 and E_e for each R_d
	Time-domain error $Tr = \{0.5 - \lvert corr\{\dot{g}(t), g_c(t)\}\rvert\}.w_t$.
Step 1b	Frequency-domain error $Sr = \{0.5 - \lvert corr\{Hg, Hc\}\rvert\}.w_s$,
Step 1c	Total Error $Tot\_err = Tr + Sr$.
Step 2	Choose five best candidates (N_{cand}) that minimizes *Tot_err*
Step 3	The transition cost is, $\delta_{i,j,k} = \{0.5 - \lvert corr\{seg_{i,j}, seg_{i-1,k}\}\rvert\}.w_{tr}$, $1 < j < N_{cand}$ $1 \le i \le M$, where M= GCIs or analysis frames
Step 4	Optimal Rd → minimize $D_{i,j} = d_{i,j} + \min\{D_{i-1,k}, \delta_{i,j,k}\}$.

**corr* is the correlation between the given variables.

The harmonic amplitudes are measured up to a frequency of *3 kHz*. The weights w_t, w_s and w_{tr} are associated with the time-domain error, frequency-domain error and transition cost, respectively. Normally, R_d value falls in the range $0.3 < R_d < 2.7$ whereas the upper range, i.e., $2.7 < R_d < 5$ signifies abduction. The R_d parameter is known to govern all the other R-parameters [**215**]. Thus, from the R_d parameter, the R-parameters are obtained as [**215**];

$$\begin{aligned} R_a &= (-1 + 4.8R_d)/100, \\ R_k &= (22.4 + 11.8R_d)/100, \\ R_g &= 1/(4 \times ((0.11R_d/(1/2 + 1.2R_k)) - R_a)/R_k). \end{aligned} \tag{6.8}$$

Using the R-parameters, the OQ can be estimated from the eq. (6.5). Thus, we consider five shape features, i.e., R_d, R_g, R_k, R_a and OQ. Figure 6.3 shows the variations of the R_d parameter for a male *Speaker A* and female *Speaker B*. For both the speakers, all the *150* natural speech utterances and *100* utterances each for *S1* VCS and *S3* SS spoof from the training set of ASV spoof 2015 challenge database are used. It is observed from Figure 6.3 that, a speaker, the R_d variations are different for natural, VCS and SS spoof. Thus, speaker-specific properties are not exactly preserved when the speech is synthesized or converted. For example, very much higher values of R_d are observed for *Speaker B* for SS than natural and VCS indicating that SS sounds more breathy. In addition, across the speakers, the R_d

variations were different, signifying that R_d captured speaker-related information both across natural and spoofed speech (which may not assist anti-spoofing).

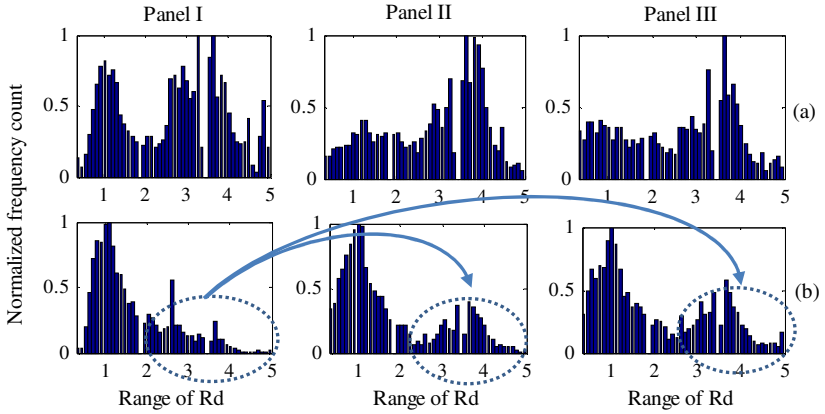


Figure 6.3: Normalized histograms of the R_d parameter for Panel I: natural speech, Panel II: vocoder-based VCS and Panel III: vocoder-based SS, corresponding to (a) *Speaker A* and (b) *Speaker B*. Dotted regions indicate the rise in R_d value for VCS and SS.

6.4 Proposed Features based on Residual Information

6.4.1 Residual in the Time Domain

Once the coarse structure $g_c(t)$ is fitted to the estimated $\dot{g}(t)$, the residual waveform is obtained as [**21**]:

$$g_r(t) = \dot{g}(t) - g_c(t). \tag{6.9}$$

The residue obtained from $\dot{g}(t)$ and $g_c(t)$ can be divided into *ripple* and *aspiration* components [**21**]. There exists a nonlinear interaction between the source and the system due to which ripple components exists in the open phase. The ripple is known to have a frequency close to that of the first formant (F_1) of the vocal tract system. It has also been shown that the ripple structure carries the speaker-specific information and, hence, is possibly a reason for the improvement in performance of speaker identification systems [**224**]- [**229**]. On the other hand, aspiration occurs due to turbulence created at the glottis when airflow passes through the partially open glottis. The amount of aspiration contributes to the quality of voice (e.g., breathy). Using the *R*-parameters estimated from eq. (6.8), the values of the timing parameters (t_o, t_e and t_c) are obtained. Considering the closed phase [0, t_o] for first glottal cycle or [t_{c-1}, t_o] for remaining glottal cycles, open phase [t_o, t_e] and the return phase [t_e, t_c], the energy (i.e., L^2 norm) of the residual $g_r(t)$ corresponding to these

regions is denoted as E_1, E_2 and E_3, respectively. For each of the glottal cycle, the energy measurements are averaged over the glottal cycle, i.e.,

$$E_1 = \frac{\widehat{E}_1}{E_{tot}}, \quad E_2 = \frac{\widehat{E}_2}{E_{tot}}, \quad E_3 = \frac{\widehat{E}_3}{E_{tot}}, \tag{6.10}$$

where $\widehat{E}_1 = \int_0^{t_o} |g_r(t)|^2 dt$ *or* $\int_{t_{c-1}}^{t_o} |g_r(t)|^2 dt$, $\widehat{E}_2 = \int_{t_o}^{t_e} |g_r(t)|^2 \, dt$, $\widehat{E}_3 = \int_{t_e}^{t_c} |g_r(t)|^2 \, dt$ and $E_{tot} = \int_{t=0}^{T_0} |\dot{g}(t)|^2 dt$

is the total energy of $\dot{g}(t)$ in a glottal cycle. For a glottal cycle, Figure 6.4 shows its corresponding estimated $\dot{g}(t)$, fitted LF-model $g_c(t)$ and its residual energy in closed phase, open phase and return phase, corresponding to E_1, E_2 and E_3, respectively. It can be observed from Figure 6.4 that during the open phase, the $\dot{g}(t)$ does not close gradually for spoofed speech as compared to natural speech due to which the ripple structure in the open phase is large for the spoofed speech (as shown by the arrows in Figure 6.4). In the regions other than open phase, the aspiration component is less for spoofed speech than in the natural speech signal.

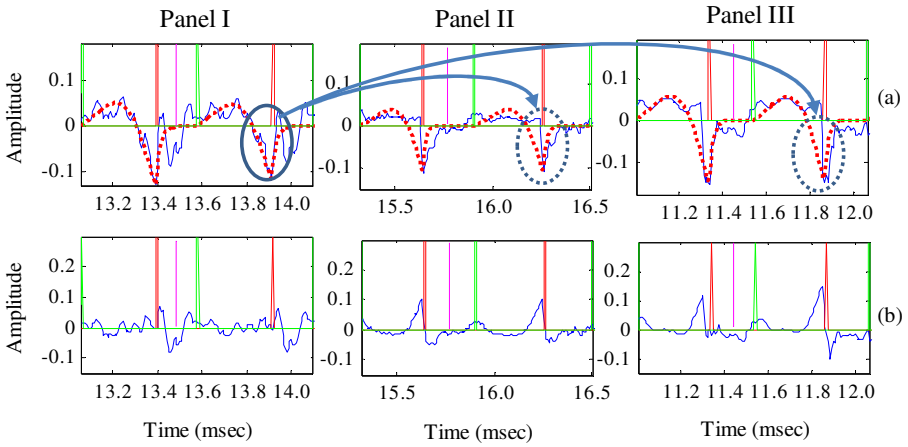


Figure 6.4: For a voiced region of speech (a) estimated $\dot{g}(t)$ and its corresponding fitted LF-model $g_c(t)$ and (b) ripple in time-domain $g_r(t)$. The glottal opening (green), GCI location (red) and glottal closing location (magenta) indicating corresponding intervals for energies E_1, E_2 and E_3, respectively, for Panel I: natural speech, Panel II: vocoder-based VCS and Panel III: vocoder-based SS (Panel III). The continuous oval in Panel I (a) indicates close match between $\dot{g}(t)$ and $g_c(t)$ whereas dotted region in Panel II (a) and III (a) indicates more deviation in fit. Adapted from [93].

As the ripple component exists in the open phase with corresponding energy E_2, we show the variations of the E_2 for *Speaker A* and *Speaker B* for all *150* natural speech utterances and all *100* utterances each for *S1* VCS and *S3* SS spoof from the training set. It is observed from Figure 6.5 that ripple energy in the open phase is

more for both VCS and SS speech and across the speakers as well. Thus, the use of ripple energy would probably has more significant contribution in SSD task.

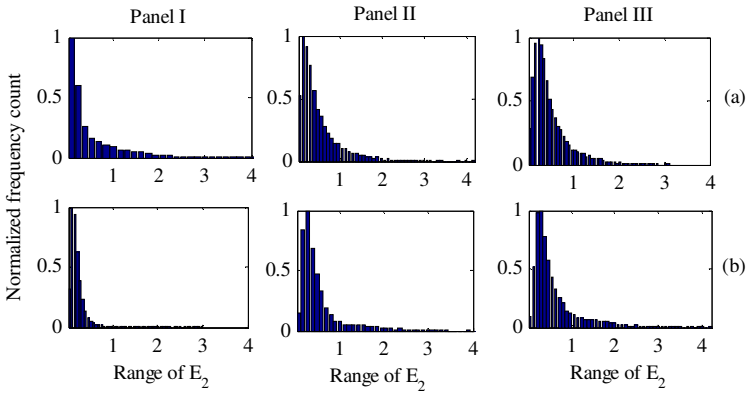


Figure 6.5: Normalized histograms of E_2 for Panel I: natural speech, Panel II: vocoder-based VCS and Panel III: vocoder-based SS corresponding to (a) *Speaker A* and (b) *Speaker B*. Adapted from [93].

6.4.2 Variation of Shape and Energy Features Across Speakers

Next, to analyze the speaker-dependency of the five shape parameters (R_d, R_g, R_k, R_a and OQ) and the three energy features (E_1, E_2 and E_3), we consider the entire training set of the ASV spoof 2015 challenge dataset (details of the dataset are given in Chapter 3, Section 3.4.1). For the *25* speakers of the training data, the mean and standard deviation of the shape and energy parameters across *150* utterances of human speech and *100* utterances for *5* spoofs is being considered here. We consider three shape parameters, R_d, R_g, and OQ (shown in Top Panel of Figure 6.6) and three energy features, E_1, E_2 and E_3 (shown in Bottom Panel of Figure 6.6). The two shape parameters, R_k, R_a are linearly related to R_d and hence, not shown here. The blue 'o' and red '*' represents the points corresponding to the mean and standard deviations of natural and spoofed speeches, respectively.

Firstly, considering the shape features, the variation across the speakers is more in natural speech than the spoofed speech for R_d and OQ parameters than R_g parameter. The R_d parameter for human speaker has spread across the normal range, $1 < R_d < 3$. However, the range for the spoofed speech was generally towards higher values of R_d, indicating more breathy voice for spoofed speech (generally, vocoder-based speech). Not much inference could be drawn from R_g parameter as the clusters of natural and spoofed speech were highly overlapping. On the other hand, the OQ parameter goes in line with R_d, i.e., the OQ had a higher mean and higher

standard deviation across all the speakers for spoofed speech as compared to human speech. Secondly, the energy features were found to represent more distinctive features for natural and spoofed speech compared to the shape features. The mean of energy E_1 in the closed phase is less for spoofed speech than the natural speech. In the case of natural speech signal, this energy is referred to as *aspiration*, which is generally noise-like. Next, as per the observations in Figure 6.4 and Figure 6.5, for E_2, the mean energy in the open phase is also more for spoofed speech than natural speech. The energy E_3 in the return phase was less varied for natural speech than for the spoofed speech. As discussed, for natural speech, the nonlinear interaction between the source and system is reflected in terms of ripple in the time-domain. However, in the case of spoofed speech, no such nonlinear interaction exists. Thus, the nature of variation in shape and energy features may result in features which may help to discriminate between the natural and spoofed speech.

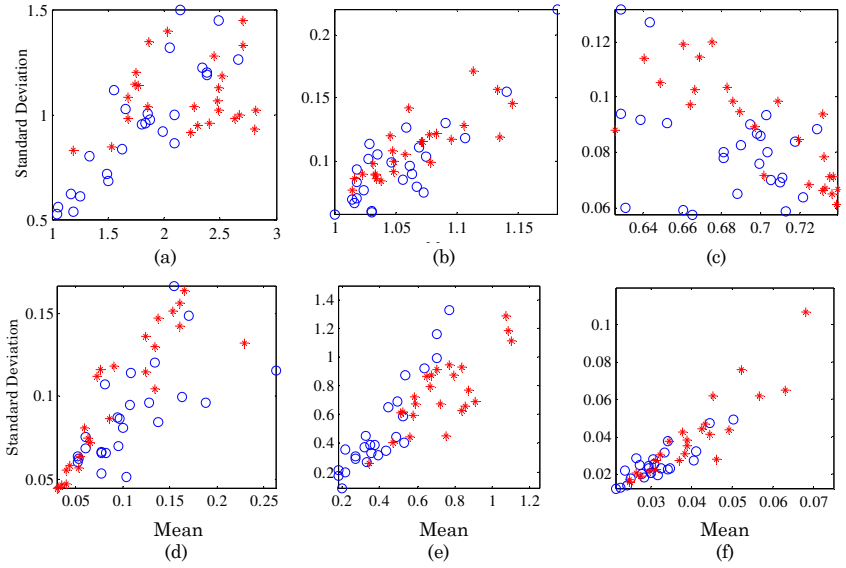


Figure 6.6: The variations in terms of mean and standard deviation. Top Panel: for three shape parameters (a) R_d, (b) R_g, and (c) OQ and Bottom Panel: for three energy features (d) E_1, (e) E_2 and (f) E_3 across all the speakers of the training dataset. Blue 'o' corresponds to speakers of the natural speech and red '*' corresponds to the spoofed speech for the same speakers. Adapted from [93].

6.4.3 Mel Representation of the Residual in Time Domain

The ripple component $g_r(t)$ is primarily due to the interaction of the excitation source with first formant (F_1) which is generally within *0-1 kHz* range. Thus, the information about the ripple is mainly embedded in the low frequency regions.

Therefore, as a secondary measure and to enhance the ripple information in the residual signal, $g_r(t)$, we consider using the Mel cepstral representation of $g_r(t)$ as features for the SSD task. In addition, it was observed that the residual, $g_r(t)$, is intelligible and, hence, the use of Mel scale that more closely mimics the human perception process for hearing (than linearly-spaced frequency bands) can be used for estimating the subband energy of this excitation source signal. Such representations of using the Mel cepstra for the excitation source, such as the $\dot{g}(t)$ [**224**] and the LP residual [**233**], has been used for speaker identification task as well. Figure 6.7 shows a speech signal, residual $g_r(t)$ and the Mel cepstral representation $g_r(t)$ for natural, VCS and SS. The VCS and SS correspond to the *S1* and *S3* algorithm of the ASV spoof 2015 challenge database, respectively. As observed for natural speech in Figure 6.7 (Panel I (c)), both low and high frequency regions are of high intensity as compared to spoofed speeches. Therefore, the ripple component and the information about aspiration component are more prominent in the natural speech as compared to the spoofed speeches shown in Figure 6.7.

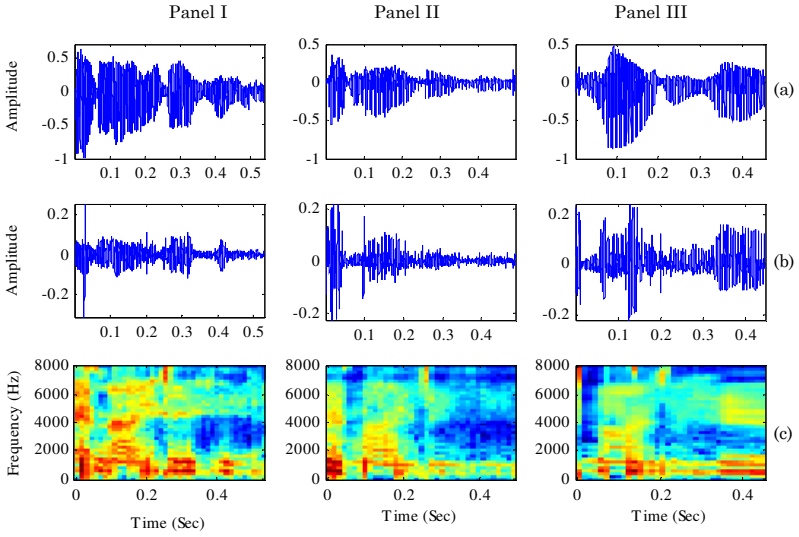


Figure 6.7: (a) Speech signal (b) residual estimated from the difference of $\dot{g}(t)$ and $g_e(t)$ and (c) the Mel representation of (b) for Panel I: natural speech, Panel II: vocoder-based VCS and Panel III: vocoder-based SS. Adapted from [93**].**

It has been observed that the higher frequency regions of speech are essential for the SSD task [**83**]. However, in the present case, the residual signal has ripple structure component with frequency around the first formant (F_1). Therefore, the Mel representation of the residual enhances this low frequency region. The high frequency regions that are speaker-specific are suppressed due to inverse filtering of

vocal tract information. Furthermore, being in the frequency-domain, this representation captures the residual information at a higher dimension than just using the average values in the closed, open and return phase.

6.4.4 Residual in the Frequency-Domain

To estimate the time-domain representation of the effect of S-F interaction, Ananthapadmanabha and Fant held the vocal tract system fixed and mapped all the nonlinear S-F interaction onto the excitation source [**224**], [**228**], [**223**]. The ripple in the time domain as a result of S-F interaction can be approximated by the following expression of $\dot{g}(t)$ [**228**]:

$$\dot{g}(t) \approx g_c(t) + f(t)e^{-0.5tB_1(t)}\cos[\int_0^t F_1(\tau)d\tau], \tag{6.11}$$

with $f(t)$ as the amplitude modulation and its multiplier reveals the first formant modulation (via *-3* dB bandwidth (B_1) and the first formant (F_1) frequency) in the frequency-domain. There exists a *duality* of ripple in time domain and formant modulation in frequency domain [**228**], [**223**]. Thus, similar to the residue in the time domain, the residue in the frequency domain has significant information of the nonlinearities due to the S-F interaction. The concept of the residue in frequency domain, (i.e., $Fr(\omega)$) is as follows,

$$\begin{aligned} Fr(\omega) &= 10\times\log\left|F\{\dot{g}(t)\}\right|^2 - 10\times\log\left|F\{g_c(t)\}\right|^2, \\ &= 10\times\log\left|\frac{F\{\dot{g}(t)\}}{F\{g_c(t)\}}\right|^2, \end{aligned} \tag{6.12}$$

where $F\{\dot{g}(t)\}$ and $F\{g_c(t)\}$ is the Fourier transform of the $\dot{g}(t)$ and $g_c(t)$, respectively. Even though we intend to compute the residual-like representation of eq. (6.9) in frequency domain, from eq. (6.12) it is clear that $Fr(\omega)$ represents the power ratio of the spectrum of $\dot{g}(t)$ and $g_c(t)$. Thus, the residue in the frequency domain will have formant modulation information at a much higher dimension than in the time domain.

Panel I and Panel II of Figure 6.8 shows the spectrograms of the estimated $\dot{g}(t)$ and fitted LF-model $g_c(t)$, respectively. The difference between these two spectra (as per eq. (6.12)) is shown in Figure 6.8 (Panel III). It is observed that the energy spreads for natural speech (Figure 6.8 (a)) and for spoofed speech (Figure 6.8 (b) and Figure 6.8 (c)) in Panel III is different for low frequency and the high frequency

regions. As in eq. (6.12), the residue of the spectrogram is also the power ratio of the spectrum of $\dot{g}(t)$ and $g_c(t)$. In the natural speech signal, as the intensity of spectrum of $g_c(t)$ is high across the entire spectrum, the intensity of residual spectrum is least across all the frequency regions as compared to spoofed speech (Panel III). To further enhance the energy variations along the frequency-axis and to use it as features, the residual spectrogram (Panel III) is divided into *36* equally spaced regions and the energy is averaged over these regions. The representation is shown in Figure 6.8 (Panel IV). This block-based energy will be further used for classification purpose in order to study the effect of low and high frequency regions for SSD task.

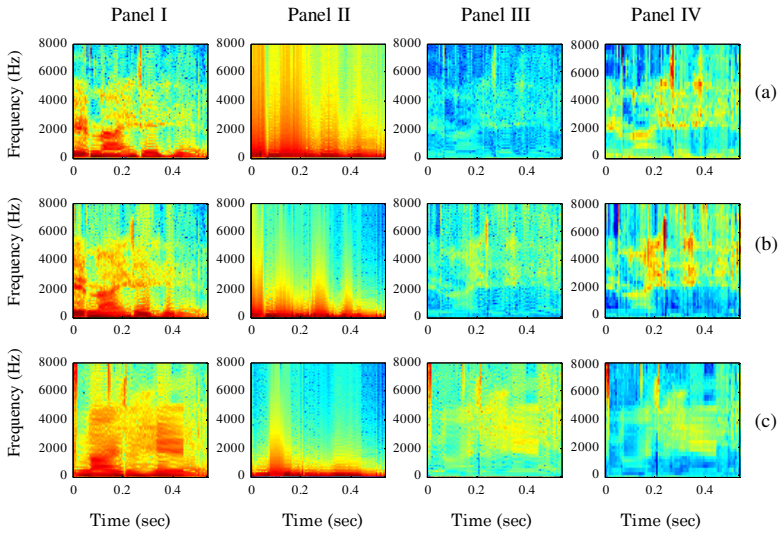


Figure 6.8: Panel I: The spectrogram of $\dot{g}(t)$, Panel II: spectrogram of the fitted LF-model $g_c(t)$, Panel III: residue in frequency-domain, i.e., difference between spectrograms of $\dot{g}(t)$ and $g_c(t)$ and Panel IV: block-based energy of residual in frequency-domain for (a) natural speech, (b) vocoder-based VCS and (c) vocoder-based SS. Adapted from [93].

Next, considering the fact that the ripple effect is mainly towards the lower frequency region, a better approach to enhance the lower frequency regions would be to use the Mel cepstral representation of the estimated $\dot{g}(t)$, the Mel cepstral representation of the fitted LF-model $g_c(t)$ and then obtain residual in the frequency-domain. Figure 6.9 shows the Mel representation of the estimated $\dot{g}(t)$ and the fitted LF-model $g_c(t)$ in Panel I and Panel II, respectively. The difference between the two representations (shown in Figure 6.9 (Panel III)) is considered as Mel warped frequency residual feature. It is observed that difference in the residue of Mel representations of $\dot{g}(t)$ and $g_c(t)$ for natural and spoofed speech differ across the utterances. In both the lower frequency region and the higher frequency region, the

energy for the natural speech was less than that of the spoofed speech. The closely spaced filters in the lower frequency region enhance the features essential for spoof detection which on fitting well for natural speech results in less energy in lower frequency regions.

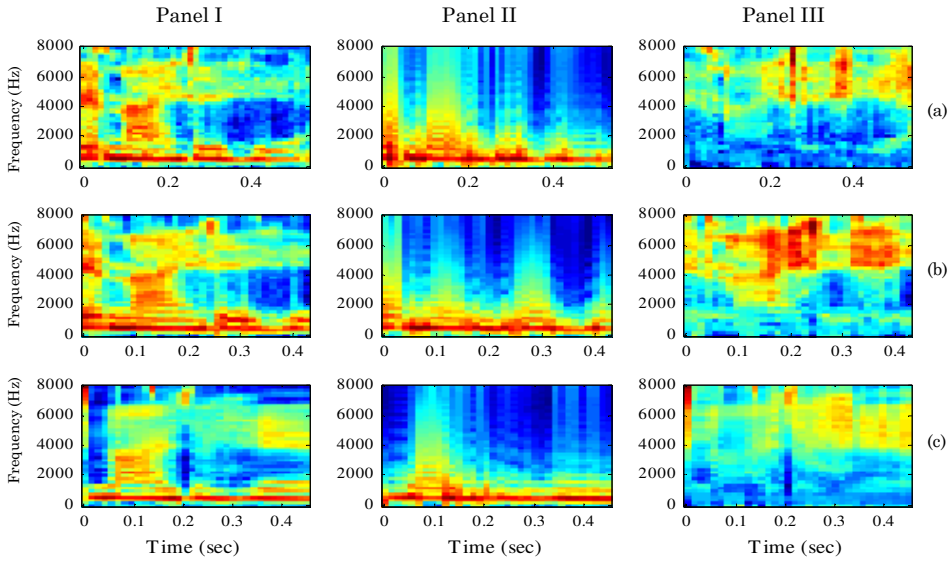


Figure 6.9: Panel I: The Mel representation of $\dot{g}(t)$, Panel II: Mel representation of the fitted LF-model $g_c(t)$ and Panel III: residue in frequency-domain, i.e., difference between Mel representations of $\dot{g}(t)$ and $g_c(t)$ for (a) natural speech, (b) vocoder-based VCS and (c) vocoder-based SS. Adapted from [93].

6.5 Experimental Results

The following Section describes in detail the parameterization carried out to develop features for the SSD task using a GMM-based classifier as described in Chapter 3. The features are designed based on several experiments and the best features are chosen for evaluation on the ASV spoof challenge data, followed by the results obtained for the signal degradation and channel mismatch conditions [**93**].

6.5.1 Parameterization

To derive the features out of the LF-model, we use the five shape-related features (R_d, R_g, R_k, R_a and OQ) and the three energy features (E_1, E_2 and E_3) in the time domain. In addition, several frequency-domain feature sets are designed based on the discussion in Section 6.4.3 and Section 6.4.4. In this context, the frequency domain residual Feature Representation ($FrFR$) is summarized in Table 6.1. The Discrete Cosine Transform (DCT) of the representations is taken to obtain static (without 0^{th} energy coefficient) and dynamic features, i.e., delta (Δ) and delta-delta

(ΔΔ). Thus, dimension (*D*) of the feature vectors, *D1*: *12-D* static features, *D2*: *24-D* (*12*s+*12*-Δ), *D3*:*36-D* (*12*s+*12*-Δ+*12*-ΔΔ) are considered. Firstly, Mel representation of the residual $g_r(t)$ is used as features, i.e., *FrFR1*. Next, the residual of the spectrogram of the estimated $\dot{g}(t)$ and the fitted LF-model $g_e(t)$ can also effectively represents the features for SSD task. Here, a *256*-point FFT is considered and as discussed earlier, the frequency-axis is divided into *36 equally* spaced regions and the energy is averaged over these regions (Section 6.4.4). The *36-D* block-based energy representation is denoted as *FrFR2*. Furthermore, the *12-D* static features of *FrFR2* obtained after DCT, which along with the *12*-Δ and *12*-ΔΔ features constitute *FrFR3*. Thereafter, the *36-D* block-based energy representation is divided into two frequency regions, called as *FrFR4*, i.e., *18-D* Low Frequency Region (LFR) for the range *0-4 kHz* and *18-D* High Frequency Region (HFR) for *4-8 kHz*. Finally, we explore the *FrFR5,* i.e., the residue of the Mel representations (instead of only FFT) as features for SSD task.

Table 6.1: The Frequency-domain residual Feature Representations (*FrFR*) for the SSD task

Feature Set	Description of Feature Set	Dimension (D)
FrFR1	Static and dynamic representation of Mel cepstral of the residue $g_r(t)$	12s+12Δ+12ΔΔ
FrFR2	Static block-based energy of residue of spectrogram of $\dot{g}(t)$ and $g_e(t)$	36-D static
FrFR3	Static and dynamic representation of *FrFR2*	12s+12Δ+12ΔΔ
FrFR4	Low frequency region (LFR) and high frequency regions (HFR) of *FrFR2*	LFR: 18-D static HFR: 18-D static
FrFR5	Static and dynamic representation of residue of Mel cepstra of $\dot{g}(t)$ and $g_e(t)$	12s+12Δ+12ΔΔ

s=static, Δ=delta, ΔΔ=delta-delta

For extracting the GCI locations, as discussed, the ZF method is used. A frame size of *30 ms* and a frame shift of *10 ms* is considered. To estimate the LF-model from the R_d search algorithm, equal weights w_t, w_s and w_{tr} are associated with the time-domain error, frequency-domain error and transition cost, respectively. From the fitted LF-model, it was observed that for the weak voiced regions, an epoch may not always be present due to which the LF-model may be ill-fitted. This resulted in the outliers in the estimated energy values in the closed, open and return phase which needs to be discarded during the training of GMMs. Thus, for time-domain features, extreme data points outside *1st* percentile to *99*th percentile are discarded. This is also done to alleviate the possible components of GMM that might model outlier distribution (especially in the case of use of a large number of mixtures in

GMM). Moreover, the presence of such outliers in features may shift the mean and variance of component Gaussians used in GMM. In addition, outliers in training or testing results in mis-classification thereby resulting in an increase in % EER. The outlier removing is done only during the training and not during the testing.

6.5.2 Results on the Development Set of ASVspoof Challenge Database

Next, as the time-domain shape and energy features are less in dimension, we experiment to find the relatively optimal number of Gaussian mixture components that would be required to model the features. Therefore, as shown in Figure 6.10, we train the GMM on various numbers of mixture components and test it using the development set. It is observed that for the shape features, no improvement in the performance was observed with the increase in the number of mixture components. While for energy features, the % EER decreased significantly with the mixture components. Even on using the shape and energy features together at feature-level, the % EER of the energy-based features alone was less.

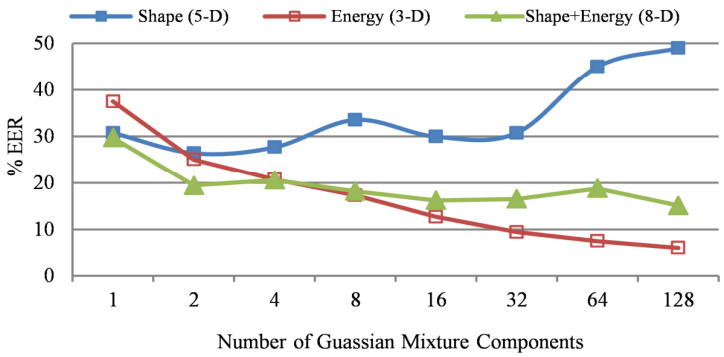


Figure 6.10: The % EER for *5-D* shape features, *3-D* energy features and the *8-D* combination of shape and the energy features at feature-level for various number of Gaussian mixture components varied from *1* to *128*.

A possible reason for shape features not performing well is that the R_d parameter is limited to the range of *0.3* to *5* and other R-parameters are derived from R_d itself. On the other hand, the energy features results due to differences between estimated and fitted model signifying the nonlinear S-F interaction. Thus, there is more possibility of capturing the differences between natural and spoofed speech. Hence, for clean speech, we consider the use of only E_1, E_2 and E_3 energy values as the time-domain features using *128* of mixture components.

Table 6.2: EER (in %) for various Frequency-domain residual Feature Representations (FrFR) on the development set

Features	FrFR1			FrFR2	FrFR3			FrFR4			FrFR5		
Dim.	D1	D2	D3	36-D	D1	D2	D3	18-D LFR	18-D HFR	Fusion	D1	D2	D3
% EER	10.06	8.17	7.80	23.65	26.85	18.93	15.27	24.02	28.96	21.90	12.30	9.67	8.81

Table 6.2 shows the % EER for the *FrFR* features. Some of the observations can be summarized as follows:

- For *FrFR1*, the EER is *10.06* % for *D1* feature vector and reduces to as low as *7.80* % using the *D3* feature vector. Thus, dynamic variations in feature trajectories are found to be effective for the SSD task.
- Next, using *FrFR2*, i.e., the *36-D* block-based energy representation yields an EER of *23.65* %.
- Using DCT of *FrFR2* to obtain *FrFR3* representation, the EER reduces from *26.85* % for *D1* feature vector to *18.9* % and *15.27* % for *D2* and *D3* feature vector, respectively.
- In order to investigate which frequency regions capture spoof-specific characteristics, the *FrFR2*, i.e., *36-D* block-based energy representation as in Fig. 8 (Panel IV) is divided into two frequency regions, called *FrFR4*, i.e., *18-D* low frequency region (LFR) for the range *0-4 kHz* and *18-D* high frequency region (HFR) for *4-8 kHz*. The EER of LFR as a feature set is found to be *24.02* % which is less than *28.96* % when the HFR are used and less than *12-D FrFR3* static features. On score-level fusion of LFR and HFR (with α_f=*0.3* as in eq. (3.6)), the EER obtained is *21.9* %, which is less than that obtained by LFR. This implies that it may be appropriate to process the residue of the spectrograms such that the LFR are enhanced more than the HFR.
- Next, to enhance the LFR, the *FrFR*5 is considered, i.e., the residue of the Mel representation of the estimated $\dot{g}(t)$ and fitted LF-model $g_c(t)$. The EER obtained with this is *12.30* % for *D1* feature vector which reduces to *9.67* % and *8.81* % with the addition of Δ and $\Delta\Delta$ features, respectively.

Hence, for the remaining set of experiments, we consider the *FrFR1* and *FrFR5* as feature vectors for SSD task. These EERs are not less than *6.09* % EER which is achieved by *3-D* time-domain energy-based features.

Score-level fusion of source features: To consider *jointly* the effect of any two feature sets, we perform a score-level fusion of the features. The fusion of two features and for three features is done as in eq. (3.6) and in eq. (3.7), respectively. It is observed in Table 6.3 that when the *FrFR1* is fused with *FrFR5* at score-level, the least EER obtained is *5.75* % using *D3* features vector. This EER is less than *FrFR1* and *FrFR5* used alone and also less than *6.09* % of the *3-D* energy features in the time-domain. The energy-based features when fused with the *D3* features vector of the *FrFR1* and *FrFR5* (with α_f=*0.2*) a significant improvement in performance is obtained resulting in *3.46* % and *3.80* % EER, respectively.

Score-level fusion of source and system features: Next, we consider the system-based MFCC feature set. The MFCC features were extracted using *28* subband Mel filters, with a frame size of *25 ms* and with *50* % overlap. On the development set, the MFCC features gave an EER of *1.6* %. With MFCC as the system-based feature set, we fuse at score-level the S-F interaction-based information, namely, energy features, *FrFR1* and *FrFR5*. As shown in Table 6.3, upon fusing the energy features with MFCC at α_f=*0.4,* the EER drops down to as low as *0.43* %. Fusing *FrFR1* and *FrFR5* at score-level with MFCC at α_f=*0.6*, the EER decreases to *0.69* % and *0.74* %,

Table 6.3: EER (in %) for score-level fusion amongst *FrFR1, FrFR5* and *3-D* energy feature sets and with system-based feature sets at various fusion factors α_f on the development set

Feature Set1	Dim	Fusion Factor (α_f)											Dim	Feature Set2
		0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**		
FrFR1	D1	10.07	9.29	8.52	8.09	7.69	**7.55**	7.75	8.09	8.89	10.41	12.30	D1	**FrFR5**
	D2	8.18	7.43	6.83	6.43	6.26	**6.23**	6.32	6.66	7.35	8.29	9.67	D2	
	D3	7.81	7.12	6.52	6.06	5.80	**5.75**	5.89	6.23	6.81	7.75	8.81	D3	
Energy	3-D	6.09	3.97	**3.46**	3.60	4.12	4.78	5.40	6.09	6.78	7.26	7.81	D3	**FrFR1**
Energy	3-D	6.09	4.29	**3.80**	3.92	4.35	5.03	5.80	6.66	7.41	8.15	8.81	D3	**FrFR5**
Energy	3-D	6.09	1.77	0.80	0.46	**0.43**	0.51	0.66	0.83	1.14	1.34	1.60	D3	**MFCC**
FrFR1	D3	7.81	4.89	3.06	1.74	1.09	0.83	**0.69**	0.80	0.92	1.32	1.60	D3	**MFCC**
FrFR5	D3	8.81	5.66	3.37	2.06	1.29	0.92	**0.74**	0.89	1.03	1.34	1.60	D3	**MFCC**

Score-level fusion is carried as per eq. (3.6)

respectively. Thus, complementary information is found to be clearly present in the S-F interaction-based features than MFCC alone for the SSD task. Moreover, the greater weightage to energy-based features indicates their relative importance than system-based features. Next, to use the system features and source features (time-domain and frequency-domain features), an equal factor of α_f=*0.5* is used. Therefore, for factors α_1=*0.4*, α_2=*0.1* and α_3=*0.5*, as in eq. (3.7), EERs of as low *0.371* % and *0.457* % is obtained as shown in Table 6.4. This % EER is very less than the best EER of *0.83* % submitted by the authors at the ASV spoof 2015 challenge using a

score-level fusion of MFCC and CFCCIF features. Hence, it is clear from the experiments that the S-F interaction residual energy features in time domain and frequency domain are highly essential to capture the vocoder-specific characteristics.

Table 6.4: EER (in %) for score-level fusion of *3-D* energy, *FrFR1* (*FrFR5*) and MFCC feature sets at selected α_f on the development set.

Score-level Fusion	% EER
α_1 . Energy + α_2 . FrFR1 + α_3 . MFCC	0.3717
α_1 . Energy + α_2 . FrFR5 + α_3 . MFCC	0.4575

Score-level fusion is carried as per eq. (3.7), α_1=0.4, α_2=0.1 and *α_3=0.5*

Dependency on spoofing algorithms: We now investigate the dependency of the features on the spoofing algorithms and spoof type (SS or VCS). Figure 6.11 shows the performance of *3-D* time-domain residual energy features, *8-D* shape plus energy features, *FrFR1* and *FrFR5* feature sets when trained individually on *S1* to *S5* and tested on the development set. As in all the cases, results are shown for known type, same type and different type of attacks. For known type of attacks, using the *3-D* time-domain energy features, the EER is very near to *< 2 %* for any training type. On an average, the % EER increases especially for VCS spoof when the shape features are used. For the frequency-domain features, the % EER decreases from *D1* to *D3* and the *FrFR1* features perform better than the *FrFR5* represenataion. On training with *S2* spoof, even known attack was not detected well with around *30 %* EER.

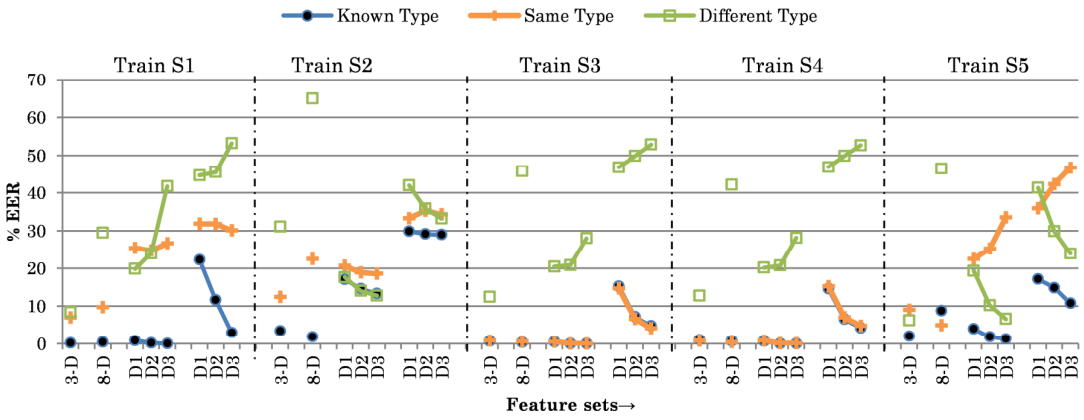


Figure 6.11: The % EER for known, same and different type of attacks when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for *3-D* energy, *8-D* shape+energy, *FrFR1* and *FrFR5* features sets and tested on the development dataset.

For the same type of attacks, the performance with SS spoof was same as for known attacks. However, for *S1* and *S2* VCS spoof, the % EER increased from *3-D* energy

features to *8-D* shape and energy features. On the other hand, the % EER was less with the shape features for training on the *S5* spoof that uses MLSA filter for speech generation process. On the contrary with the frequency-domain features, the detection using *S1* and *S2* in the training decreases the % EER from *D1* to *D3* feature vector while the % EER increases on training with the *S5* spoof alone. For different type of attacks, the *3-D* energy features gave the least % EER (except *S2* spoof) amongst all feature set. The performance of frequency-domain features for SS speech is similar to previous features where SS cannot detect VCS spoof. However, the *S1* VCS when used in training did not detect SS and the % EER increased from *D1* to *D3* feature vector. Thus, to extract significant spoof-specific information from the S-F features, training is essential using both the spoof types (SS and VCS).

Discussion on the DET curves: The DET curves for the energy features, *FrFR1* features, *FrFR5* features and their score-level is shown in Figure 6.12. It is observed that without fusion, the *FrFR5* has the relatively highest % EER and the energy-based features have the least % EER. However, the FAR of the time-domain energy features were far more than *FrFR1* and *FrFR5* for FRR less than *2* %. Both the *FrFR1* and *FrFR5* features had more % FRR than the time-domain energy-based features. Thus, it is a feasible option to fuse the energy-based features in time-domain and the frequency-domain representations for better performance. The score-level fusion of time-domain energy features with *FrFR1* and *FrFR5* features decreased the % EER as well as the % FRR significantly. It was observed that after fusion, the % FAR did not reduce much for less than *0.5* % FRR.

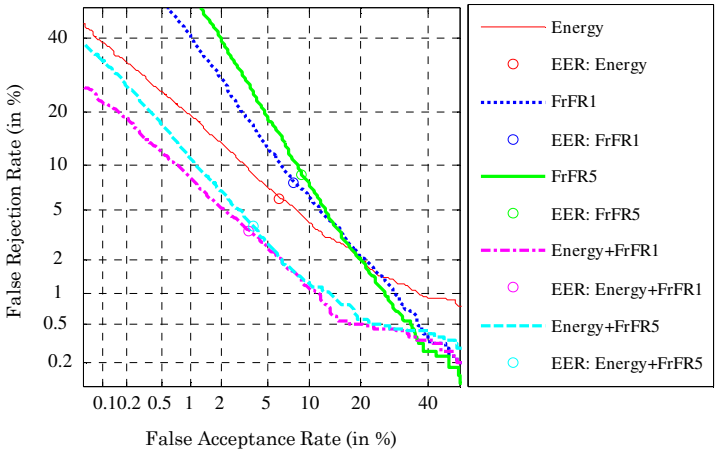


Figure 6.12: DET curve on the development set for *3-D* time-domain energy features (red), *FrFR1* (blue), *FrFR5* (green), score-level fusion of *3-D* energy features with *FrFR1* (magenta) and *FrFR5* (cyan) at α_f=*0.2*.

6.5.3 Results on the Evaluation Set of ASVspoof Challenge Database

In the realistic scenarios, the type of spoof (SS or VCS) or the spoofing algorithm will not be known. Thus, the performance is studied by testing with all *S1-S10* spoofs of the evaluation set. As analyzed from the development set, a score-level fusion of source- and system-based features gave significantly lower % EER than the features used individually. Considering the features and the fusion factors obtained from the development set, the overall % EER on the evaluation set and the % EER of the individual *S1* to *S10* attacks are shown in Table 6.5.

Table 6.5: EER (in %) in terms of individual attacks, average known attacks, average unknown attacks, average with and without *S10* spoof for time-domain energy features, *FrFR1* and *FrFR5* features, score-level fusion of time-domain energy features, *FrFR1* and *FrFR5* features and MFCC at selected α_f on the evaluation set

Feature sets	Dim.	Individual Attacks										Average		
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn	w/oS10	Avg
Energy	3-D	0.212	6.902	0.130	0.163	1.777	4.005	21.179	2.141	7.853	86.429	1.837	4.930	13.070
Shape + Energy	8-D	4.174	17.050	9.277	9.120	14.590	21.760	26.908	3.342	10.650	81.087	10.840	12.990	19.790
FrFR1	D1	7.033	14.543	1.832	1.897	8.462	13.120	5.152	1.348	3.946	84.304	6.752	6.370	14.160
	D2	5.457	13.560	0.630	0.761	6.848	11.543	2.576	1.234	3.668	84.212	5.452	5.142	13.050
	D3	5.772	15.212	0.842	0.761	5.783	10.174	2.098	1.755	3.853	82.750	5.672	5.137	12.900
FrFR5	D1	4.136	21.853	6.348	6.848	8.745	10.082	27.451	12.473	7.397	77.120	9.586	11.700	18.250
	D2	1.951	22.766	3.016	3.043	7.826	8.299	16.402	9.967	9.402	76.397	7.721	9.186	15.910
	D3	0.962	21.359	2.283	2.337	7.087	7.842	8.375	8.560	7.353	76.761	6.805	7.351	14.290
Energy +FrFR1	D3+3-D	0.038	3.092	0.011	0.033	0.505	1.815	1.332	0.212	1.321	88.375	0.736	0.929	9.673
Energy +FrFR5	D3+3-D	0.054	4.815	0.043	0.065	0.690	1.511	5.272	0.826	2.457	86.761	1.134	1.748	10.250
MFCC	D3	0.005	0.995	0.000	0.000	0.832	0.902	0.054	0.000	0.082	39.723	0.366	0.319	4.259
Energy +MFCC	3-D+D3	0.000	0.141	0.000	0.000	0.027	0.038	0.000	0.000	0.005	47.913	0.034	0.024	4.813
FrFR1 +MFCC	D3+D3	0.000	0.196	0.000	0.000	0.125	0.304	0.016	0.000	0.005	47.402	0.064	0.072	4.805
FrFR5 +MFCC	D3+D3	0.000	0.571	0.000	0.000	0.321	0.272	0.016	0.000	0.022	44.011	0.178	0.133	4.521
Energy +FrFR1 +MFCC	3-D+D3+D3	0.000	0.065	0.000	0.000	0.027	0.060	0.000	0.000	0.000	42.261	0.018	0.017	4.241
Energy +FrFR5 +MFCC	3-D+D3+D3	0.000	0.174	0.000	0.000	0.049	0.043	0.000	0.000	0.005	40.973	0.045	0.030	4.124

Score-level fusion is carried as per eq. (3.6) and eq. (3.7), Kn=known, w/o S10=Average without S10, Avg. = Average of S1-S10

Considering the EER for the known attacks, for the shape and energy features used together at feature-level (*8-D*), the % EER is very high and hence, not considered for fusion. With the *3-D* time-domain energy features, the EER is *1.84* % which is significantly less. The *FrFR1* and *FrFR5* with *D3* feature vector give *5.672* % and *6.805* % EER. A score-level fusion of energy features with *FrFR1* and *FrFR5*

at $\alpha_f=0.2$ gives *0.736* % and *1.134* % EER, respectively. Secondly, considering the system-based MFCC features, the EER is found to be around *0.36* % for known attacks. On combining the energy and MFCC features at $\alpha_f=0.4$, the % EER reduced *10* times compared to MFCC to achieve an EER of *0.034* %. Similar observations were observed when *FrFR1* and *FrFR5* features were fused at score-level with MFCC at $\alpha_f=0.6$. Thus, use of S-F interaction-based features for SSD can be justified.

The evaluation set consists of both vocoder-dependent and vocoder-independent speech. Hence, the overall % EER is obtained without *S10* (only vocoder-based (*S1-S9*)) and with *S10*. The % EER of vocoder-based spoofs (*S1-S9*) is almost similar to that of the known case. That is, the vocoder-based attacks gave a significantly low EER of *0.017* % on the score-level fusion of the energy, *FrFR1* and MFCC features with factors of $\alpha_1=0.4$, $\alpha_2=0.1$ and $\alpha_3=0.5$. An example of the effectiveness of score-level fusion is the *S7* spoof for which the time-domain energy features gives an EER of *21.18* %. However, when fused at score-level with *FrFR1* and *FrFR5* features, the EER reduced to *1.332* % and *5.272* %, respectively. The EER for *S7* spoof decreases to *0.00 %* on the score-level fusion of energy and MFCC features.

Considering the % EER of the *S10* spoof, it is observed that the relatively best % EER was obtained using the MFCC features. It is observed that the S-F interaction-based features do not contribute significantly when used alone or with the MFCC feature set. Slightly lower EER of *4.124* % than MFCC alone was obtained by fusing energy, *FrFR5* and MFCC features with fusion factors of $\alpha_1=0.4$, $\alpha_2=0.1$ and $\alpha_3=0.5$, respectively. However, this decrement is not very significant. Thus, the S-F interaction features do not contribute in detecting the spoofed speech due to concatenative speech synthesis, where, in principle, natural speech sound units are joined and hence, create a lot more confusion during classification of natural and spoofed speech. This is due to the fact that the residual features are characteristics of the natural speech which are still preserved in the *S10* spoof due to the direct concatenation of the speech sound units. Thus, the residual features in time and frequency-domain may not prove to be much effective.

Dependency on spoofing algorithms: On the evaluation set, to check the spoof dependency of the features, we carry the same evaluation as on the development set. The analysis of same and different type is almost similar to that obtained on the development set. Considering the performance of the features on the *S10* spoof,

Figure 6.13 shows that as compared to the results for vocoder-based spoofs, for *S10* the % EER decreases with the use of *8-D* shape and energy-based features. This may

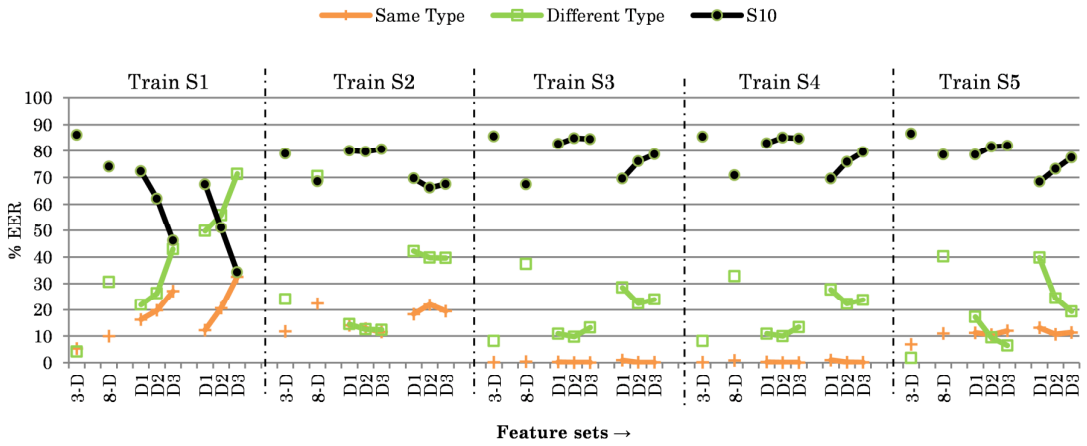


Figure 6.13: The % EER for the same type, different type and *S10* attack when trained with individual spoofs *S1*, *S2*, *S3*, *S4* and *S5* for *3-D* energy, *8-D* shape+energy, *FrFR1* and *FrFR5* features sets and tested on the evaluation dataset.

be due to the fact that in a particular speech utterance, for a particular speaker, the variation of shape parameter depends only on the type of utterance. On the other hand, in unit-selection speech, the speech sound units are concatenated and which may again vary in voice quality due to recordings from various sessions. Thus, more variations in voice quality exist for *S10* spoof which helps in reducing the % EER slightly than the energy features used alone. This may possibly be the reason for the decrease in % EER of *S10* in Table 6.5 when the shape and energy features used jointly by their feature-level fusion. Even though the % EER has decreased, the improvement is not very significant as only one particular aspect of voice quality difference in natural and unit-selection speech is observed at a very low dimension.

6.5.3.1 Comparison with Other Features

The performance of the S-F interaction features is compared with the previous work on using cochlear filter-based CFCC, CFCCIF, CFCCIFS and feature sets as well as other excitation source features such as, F_0 and *SoE* features and prediction residual-based features. The SBAE features are not discussed as its performance with the MFCC features are not better than the cochlear filters. These features were also fused at score-level with MFCC. A brief description of the features and their relative performance as compared to S-F interaction features is given as follows.

Cochlear-based features: The CFCC feature sets are based on using the auditory filterbanks as compared to triangular filterbanks in MFCC. In addition, the envelope at the output of the cochlear subband is combined with the average subband IF information. Moreover, to capture transient information or the variation across the frames, the derivative operation is used. These features are known as CFCCIF. In addition, the use of symmetric difference to estimate variations of subband energy representation (i.e., CFCCIFS) has shown to give better performance [**91**]. As shown in Table 6.6, the cochlear-based features, when combined with MFCC using score-level (α_f=*0.2* as per eq. (3.6)) gave a very low average % EER of *1.44* %. However, for the vocoder-based *S1-S9* spoofs, the EER was *0.16* % which is almost ten times more than with the energy and *FrFR1* features when fused with MFCC at score-level.

Table 6.6: % EER on testing with the evaluation set for cochlear-based CFCC, CFCCIF, CFCCIFS features and source-based features when fused at score-level fusion with MFCC feature set

Feature Set	Dim.	Individual Spoofing Attacks (% EER)										Overall % EER		
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn.	S1-S9	S1-10
CFCC+MFCC	**D3+D3**	0.01	0.74	0.00	0.00	1.33	0.68	0.07	0.00	0.12	13.08	0.41	0.32	1.60
CFCCIF+MFCC	**D3+D3**	0.00	0.36	0.00	0.00	0.97	0.50	0.04	0.08	0.04	16.72	0.26	0.22	1.87
CFCCIFS+MFCC	**D3+D3**	0.00	0.24	0.00	0.00	0.72	0.31	0.03	0.09	0.03	13.03	0.19	0.16	1.44
(F0-SoEs) +MFCC	**12-D+D3**	0.00	0.72	0.00	0.00	0.19	0.30	0.02	0.00	0.03	34.47	0.18	0.14	3.57
Best(M1+M2) +MFCC	**24-D+D3**	0.00	0.04	0.00	0.00	0.02	0.02	0.01	0.00	0.01	51.11	0.01	0.01	5.12

F_0 and SoE features: These features are based on the fact that when the vocal folds vibrate, there exists a correlation between the F_0 contour and *SoE* at the glottal excitation source (*SoE1*) and at the speech signal (*SoE2*), which is found to be more for natural speech than machine-generated speech [**86**]. Moreover, as natural speech has more variations, the dynamics of the F_0, *SoE1* and *SoE2* features are also considered by taking their derivative up to 3^{rd} order. These features when combined with MFCC using score-level (α_f=*0.8* as per eq. (3.6)) perform slightly better than cochlear-based features for *S1-S9* spoofs. However, even in this case for *S1-S9* spoofs, the % EER of proposed S-F features is almost ten times better.

LP-LTP and LP-NLP features: Here, the LP, LTP and NLP features are explored based on the idea that the spoofed speech is too easy to predict if a simplified acoustic model generates it and it is too difficult to predict if there are artifacts present in the speech signal [**76**], [**90**]. Hence, the score-level fusion of LP-LTP (M1)

and LP-NLP (M2) combination at $\alpha_f = 0.4$, when further combined with MFCC at score-level ($\alpha_f = 0.1$ as per eq. (3.6)) provided discriminative or complementary features especially for *S1-S9* spoofs. The performance for only vocoder-based spoofs is slightly better than the S-F interaction features. However, the EER for prediction-based features is high for *S10*; as a result the average % EER is more than the proposed S-F interaction features.

Fusion of S-F interaction features with CFCCIFS feature set: From Table 6.5 and 6.6, it is observed that S-F interaction features work well in detecting the vocoder-based speech (*S1-S9*). On the other hand, the previous work of using envelope and IF information jointly (i.e., CFCCIF and CFCCIFS features) gave reduced %EER for *S10* spoof as compared to S-F interaction or other source-based features. Therefore, we attempt to combine the benefits from both of these two features as shown in Table 6.7. Firstly, for combination of time-domain energy features, FrFR1/FrFR5 and CFCCIFS features, the fusion factors are $\alpha_1 = 0.4$, $\alpha_2 = 0.1$ and $\alpha_3 = 0.5$. It is observed that with the addition of CFCCIFS features, the EER of *S10* decreases to around ~*15-16%*. On increasing the contribution of CFCCIFS features (i.e., $\alpha_1 = 0.2$, $\alpha_2 = 0.1$ and $\alpha_3 = 0.7$), the average EER decreases to around ~*1.4* % where the EER of *S10* decreases to *13.08* and *12.83* on using *FrFR1* and *FrFR5* features, respectively. The performance of *S1-S9* slightly degrades as compared to S-F interaction and MFCC features, however, the performance of *S1-S9* is better than CFCCIFS +MFCC system. It is to be noted that the use of attack-independent threshold makes the detection of *S10* even more difficult and hence responsible for large average EER.

Table 6.7: EER on Testing with the evaluation set for score-level fusion of time-domain energy features, *FrFR1* and *FrFR5* features with CFCCIFS feature set

Feature Set	FF	Individual Spoofing Attacks (% EER)										Overall % EER		
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Kn.	S1-S9	S1-10
Energy+FrFR1+CFCCIFS	A1	0.00	0.01	0.00	0.00	0.03	0.06	0.01	0.14	0.01	16.14	0.01	0.029	1.640
Energy+FrFR5+CFCCIFS	A1	0.00	0.01	0.00	0.00	0.07	0.07	0.00	0.20	0.03	15.18	0.01	0.043	1.557
Energy+FrFR1+CFCCIFS	A2	0.01	0.08	0.00	0.00	0.26	0.14	0.01	0.33	0.04	13.08	0.07	0.097	1.396
Energy+FrFR5+CFCCIFS	A2	0.01	0.14	0.00	0.00	0.35	0.12	0.04	0.45	0.06	**12.84**	0.10	0.130	**1.401**

Fusion Factors (FF) A1: $\alpha_1 = 0.4$, $\alpha_2 = 0.1$ and *$\alpha_3 = 0.5$, A2: $\alpha_1 = 0.2$, $\alpha_2 = 0.1$* and *$\alpha_3 = 0.7$*

Results on signal degradation conditions: Amongst the several approaches used for SSD task, the results are mostly evaluated in the presence of clean conditions. Very recently, research had been directed towards evaluating the performance of

countermeasures in the presence of noisy environments. In [**94**], a preliminary investigation of spoofing detection under additive noisy conditions had been performed. This work also describes an initial noisy database developed by artificially adding background noises at different SNR levels. This work shows that for a model trained on clean data, the system performance degrades significantly when tested on noisy speech. It was observed that the system performance differs with the types of noises. In [**96**], on the similar grounds, several countermeasures were found to fail at relatively high SNRs and did not generalize well for the SSD task even with speech enhancement algorithms. In this study, we consider evaluating the performance of the proposed S-F interaction features for additive white noise, babble noise and car noise at various SNR levels, namely, *10* dB, *5* dB and *0* dB. The performance evaluation for the various features in signal degradation conditions is shown in Table 6.8.

Table 6.8: % EER of the source and system features for different feature sets on the evaluation set in the presence of additive white noise, babble noise and car noise at various SNR levels

Feature Sets		Energy		Shape +Energy		FrFR1						FrFR5						MFCC					
Dim.		3-D		8-D		D1		D2		D3		D1		D2		D3		D1		D2		D3	
% EER → / SNR (dB) ↓		S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.	S1-S9	Avg.
Clean		4.93	13.1	12.9	19.8	6.37	14.2	5.14	13.1	5.13	12.9	11.7	18.3	9.19	15.9	7.35	14.3	1.48	5.49	0.56	5.49	0.32	4.25
White	10	3.6	11.5	13.1	19.7	23.4	28.1	12.9	19.6	12.0	18.8	24.0	28.0	25.5	28.8	26.9	30.2	39.8	42.5	40.3	43.6	38.4	41.5
	5	8.3	13.6	14.5	19.2	29.4	33.8	21.1	27.4	18.6	24.9	34.7	36.9	30.9	33.1	32.9	35.2	41.6	44.1	48.1	51.5	40.9	43.2
	0	31.3	31.3	28.4	29.6	39.2	41.3	44.3	46.8	42.5	45.2	53.6	53.5	49.0	49.5	46.4	47.0	42.0	44.1	50.5	53.8	43.1	44.1
Average		14.4	18.8	18.7	22.8	30.7	34.4	26.1	31.3	24.4	29.6	37.4	39.5	35.1	37.1	35.4	37.5	41.1	43.6	46.3	49.6	40.8	42.9
Babble	10	8.7	16.8	14.3	20.6	16.9	23.4	14.2	21.3	16.0	22.8	22.0	27.1	22.0	27.5	20.2	25.7	45.8	50.2	35.7	40.9	30.8	36.3
	5	12.9	20.1	19.5	24.5	25.9	30.8	22.6	28.2	25.3	30.3	29.2	33.2	29.9	34.3	27.8	32.0	45.6	49.6	45.0	49.0	40.2	44.0
	0	27.5	32.1	27.8	31.7	33.4	36.5	33.8	36.8	36.5	39.1	29.8	32.8	25.6	29.6	24.9	28.6	42.9	46.0	48.7	51.7	44.5	46.8
Average		16.4	23.0	20.5	25.6	25.4	30.2	23.5	28.8	25.9	30.7	27.0	31.0	25.8	30.5	24.3	28.8	44.8	48.6	43.1	47.2	38.5	42.4
Car	10	2.7	11.6	15.0	22.0	7.8	15.6	7.3	15.4	7.4	15.3	17.2	22.8	14.1	20.3	12.6	19.0	20.0	26.7	9.6	16.1	15.1	22.8
	5	4.4	13.1	19.6	26.0	11.1	18.7	13.2	21.2	13.6	21.5	19.3	24.5	17.7	23.5	16.8	22.7	27.1	33.4	12.6	18.9	21.6	29.0
	0	13.6	20.7	22.3	27.9	17.1	23.6	19.3	26.0	20.7	27.5	29.8	32.8	25.6	29.6	24.9	28.6	34.6	40.1	15.9	21.8	28.4	35.2
Average		6.9	15.1	19.0	25.3	12.0	19.3	13.3	20.9	13.9	21.4	22.1	26.7	19.1	24.5	18.1	23.4	27.2	33.4	12.7	18.9	21.7	29.0

The performance is shown in terms of % EER for vocoder-based (*S1-S9*) spoofs and overall % EER (*S1-S10*). For white noise, it is observed that the energy features gave almost equal average % EER in clean and in the presence of *10* dB and *5* dB noise. A similar case was observed with that of the shape and energy features fused at feature-level. The % EER of the shape and energy features, when used jointly, was more than the energy features till *5* dB. However, for *0* dB SNR, the use of shape

and energy features gave the relatively better performance of *29.6* % EER. For these features, the spoof-specific information was preserved at severe signal degradation conditions as well. The % EER of the *8-D* shape and energy features seems to decrease with the degradation. However, at *10* dB, the % EER increases for *S1-S9* and decreases for *S10* spoof and hence, the overall % EER shown in Table 6.6 decreases slightly. On the other hand, the frequency-domain features were severely affected by noise. Especially for MFCC feature set, at *0* dB SNR, the performance degrades to around *44* % with the *D3* feature vector. The *FrFR1* and *FrFR5* features were found to perform better than MFCC till *5* dB SNR noise.

For babble noise, the % EER for all features increases even for *10* dB SNR. The % EER is least for *3-D* energy features and is maximum for MFCC feature set. For *0* dB SNR, the least average EER is *28.6* % obtained using *FrFR5* feature set. The *FrFR1*and *FrFR5* representations, i.e., the S-F interaction cues in the frequency domain were able to better classify natural *vs.* spoofed speech in the presence of babble noise as compared to white noise. However, on an average of all SNRs, the *3-D* energy features perform the best amongst all the features. Next, for car noise, the % EER for all the features (except *3-D* energy features) increased at *10* dB SNR. However, the performance degradation was less as compared to white and babble noise. Interestingly, at *10* dB SNR using *3-D* energy features, the % EER improved both for vocoder-based cases (*S1-S9*) as well as for the average % EER. This is because the % EER of vocoder-based *S7* spoof decreased about *10* times in the presence of *10* dB car noise. This was also observed for *10* dB white noise as well where the detection for *S7* and *S10* was improved. Similar observations were found in [**94**], where the performance on *S10* improved in the presence of reverberation noise. It was observed that with the temporal filtering of reverberation, the discontinuity in *S10* spoof could become more obvious. However, much needs to be explored about the improved performance even in the presence of noise.

In recent works, where white noise is considered [**94**]- [**96**], it is observed that the % EER at *0* dB is as high as *40* % obtained by fusing several features with each other. In such multiple fusions of features, it is difficult to conclude as to why a particular feature performs well for a particular noise. In the present case, the *8-D* shape and energy features gave better performance even at *0* dB SNR. This is because, in the present approach, the excitation source features in time domain are

obtained by inverse filtering from the speech signal the high frequency resonances corresponding to the vocal tract system. Thus, the high frequency noise is also filtered out. Therefore, the shape parameters along with the energy-based features help in maintaining the performance of the SSD systems much better as compared to the MFCC features that contain much broader spectra.

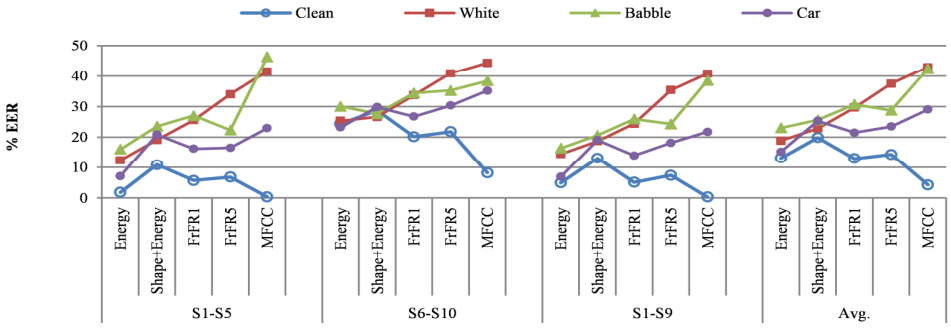


Figure 6.14: The % EER for known attacks (*S1-S5*), unknown attacks (*S6-S10*), vocoder-based spoofs (*S1-S9*) and average EER (*S1-S10*) averaged across the various SNR levels for *3-D* energy features, *8-D* shape and energy features, FrFR1, FrFR5 and MFCC feature sets for *D3* dimension.

Figure 6.14 shows the % EER for known attacks (*S1-S5*), unknown attacks (*S6-S10*), vocoder-based spoofs (*S1-S9*) and average EER (*S1-S10*) (averaged across the various SNR levels) for *3-D* energy features, *8-D* shape and energy features, *FrFR1*, *FrFR5* and MFCC feature sets for the *D3* dimension. It can be observed that MFCC is highly sensitive to any type of signal degradation conditions as compared to the frequency-domain features that use S-F cues. The simple approach of residual using energy-based features proves to be effective in the presence of noise as well without significant performance degradation. It is also observed that, on average, white and babble noise were more severe as compared to car noise and at low SNR values, the white noise affects the performance more than the babble noise.

6.5.4 Results on the Blizzard Challenge 2012 Database

The results of the S-F interaction features with *3-D* energy features, *8-D* shape and energy features, *FrFR1* and *FrFR5* representation are shown in Table 6.9. As observed, the problem of channel mismatch cannot be generalized based on the performance of countermeasures on the ASV spoof challenge database. Generally, the performance of MFCC decreases with the addition of Δ and $\Delta\Delta$ features to the static features. However, in this case, the % EER for MFCC does not always decrease

with the addition of dynamic features as compared to the gradual decrease in % EER by the *FrFR1* and *FrFR5* features. Thus, MFCC with *D3* feature vector cannot be considered optimum in all the cases. On the ASV spoof challenge data, the shape features did not contribute in decreasing the % EER. However, for completely unknown data, the synthetic speech recordings by several systems were detected by *8-D* shape and energy features. The results are much better than MFCC features.

For the shape and energy features, the % EER is very less for statistical-based synthetic speech as compared to USS-based, hybrid or diphone-based synthetic speech. Considering an average representation across USS, statistical, hybrid and diphone-based systems, amongst all the systems, the USS-based systems were found to be difficult to detect in the SSD task. In the case of statistical systems, the shape and energy-based features gave very less % EER. For hybrid and diphone systems, the performance is similar to that of the USS-based speech recordings (except system *G* that gave *5* % EER with the *8-D* shape and energy feature set). The *FrFR5* features detected diphone systems with least *22* % EER. On the whole, for the channel mismatch case, the performance of shape and energy features were found to be better than the other features. The performance of the features, in this case, is highly dependent on the type of attacks. Therefore, there is a need to look into the type of training strategy used such that the channel variability can be handled.

Table 6.9: EER (in %) for *3-D* energy, *8-D* shape+energy, *FrFR1* and *FrFR5* features sets on training with the ASV spoof data and testing with the Blizzard Challenge 2012 database

2012 Blizzard	System	Energy	Shape+Energy	*FrFR1*			*FrFR5*			MFCC		
		3-D	8-D	D1	D2	D3	D1	D2	D3	D1	D2	D3
USS	**B**	57	**47**	55	60	56	59	57	56	98	77	67
Hybrid	**C**	62	53	48	53	**46**	54	52	61	40	46	47
Hybrid	**D***	39	24	90	85	80	37	20	**15**	65	66	42
HMM	**E***	6	**0**	48	43	33	15	8	2	44	82	61
USS	**F**	52	**37**	79	83	77	52	54	61	22	24	15
USS	**G**	17	**5**	60	65	60	28	22	18	8	29	27
HMM	**H**	**1**	3	5	5	**1**	12	5	4	12	38	3
USS	**I**	53	**48**	55	59	54	61	58	55	98	97	69
Diphone	**J***	37	38	79	67	67	25	24	**22**	64	69	69
HMM	**K***	9	**0**	55	65	54	11	8	1	92	67	73

* systems with lower MOS from *1* ≤ *2*

6.5.5 Results on the Blizzard Challenge 2014 Database

The results of the source-system interaction features are given in Table 6.10 and Table 6.11 for Gujarati and Hindi dataset, respectively. For the Gujarati dataset, the HMM-based synthetic speech was detected easily by the energy features alone

(except that for system *H*). The USS-based *G* system was not detected either by time-domain features or frequency-domain *FrFR1* and *FrFR5* features. However, the MFCC features detected it with *34* % EER. The HMM-DNN-based *F* system was not detected well by S-F interaction-based features. However, the MFCC feature set could detect system *F* with *6* % EER. The synthesized utterances from the HMM-based systems in Hindi language were detected well using *3-D* energy, *FrFR1* and *FrFR5* feature sets. On the other hand, the % EER increased with the use of *8-D* shape and energy features for all the systems. The Hybrid and USS systems were difficult to detect with EER around *60 % -70 %*. The HMM-DNN system was detected by MFCC with least EER of *7* % with the *D2* feature vector. Thus, on an average, the *3-D* energy features generalized well to detect the unknown HMM-based speeches both for Blizzard 2012 and Blizzard 2014 datasets.

Table 6.10: EER (in %) for *3-D* energy, *8-D* shape+energy, *FrFR1*, *FrFR5* and MFCC features sets on training with the ASV spoof data and testing with Blizzard Challenge 2014 database for the Gujarati language

Blizzard 2014	Gujarati Systems	Energy	Shape+Energy	*FrFR1*			*FrFR5*			MFCC		
		3-D	8-D	D1	D2	D3	D1	D2	D3	D1	D2	D3
HMM	**C**	**1**	4	2	2	3	10	10	5	37	13	4
HMM	**D**	**0**	30	9	3	0	7	5	1	67	6	1
HMM	**E**	**0**	0	0	1	0	9	8	4	13	21	4
HMM-DNN	**F**	27	22	89	91	95	23	30	28	75	23	**6**
USS	**G**	**34**	66	76	87	87	41	70	71	67	48	**34**
HMM	**H**	46	72	83	83	84	40	50	47	55	24	**24**

* wavefiles for baseline system B and system I are not available

Table 6.11: EER (in %) for *3-D* energy, *8-D* shape+energy, *FrFR1*, *FrFR5* and MFCC features sets on training with the ASV spoof data and testing with Blizzard Challenge 2014 database for the Hindi language

Blizzard 2014	Hindi Systems	Energy	Shape+Energy	FrFR1			FrFR5			MFCC		
		3-D	8-D	D1	D2	D3	D1	D2	D3	D1	D2	D3
HMM	**B***	**0**	7	2	**0**	**0**	10	3	2	2	5	14
HMM	**C**	**1**	11	1	**0**	**0**	26	10	6	1	4	6
Hybrid	**D**	76	79	68	72	65	78	73	74	68	**25**	62
HMM	**E**	2	6	**0**	**0**	**0**	13	12	1	**0**	5	7
HMM-DNN	**F**	35	45	59	64	56	47	29	31	59	**7**	19
USS	**G**	38	65	62	75	63	70	66	73	62	52	52
HMM	**H***	**0**	3	2	1	1	6	8	3	2	11	30
HMM	**K**	**0**	14	2	1	**0**	8	2	2	2	8	32

* systems with lower MOS from *1* ≤ *2* (wavefiles for system I are not available)

6.6 Chapter Summary

This study presented the use of the features motivated from the natural human speech production mechanism. That is, each time the vocal folds open and close, there exists a nonlinear source and system interaction which can be estimated from

the residual component $g_r(t)$ in time-domain and frequency-domain. We presented the significance of R_d shape parameter of the LF-model in interpreting the characteristics or quality of speech. However, not much can always be inferred about the naturalness of speech due to the fact that the vocoded speech is a subset of the natural speech signal. In addition to the shape parameter, the residual energy in the closed, open and return phase is considered. The extensive analysis in this study indicates that the energy in the open phase specifically aids in the spoof detection task. The S-F interaction features represent information of the voice excitation source at a lower frequency region than the actual speech signal. Thus, this is indeed a promising approach as these features are likely to be robust to signal degradation conditions. Furthermore, the testing results on a completely unrelated database such as the Blizzard Challenge showed that the time-domain S-F interaction features perform very well. Spoof detection in signal degradation and channel mismatch conditions is an important research issue and needs further investigations as the features should generalize in terms of better performance for the clean speech as well. In the next Chapter, we summarize the performance of the several features discussed in this thesis and evaluate them based on various factors taking into consideration the present problem of spoof detection.

Chapter 7.
Summary and Conclusions

7.1 Summary of the Work

The study presented in this thesis tries to address the problem of spoof detection by exploring the three basic aspects of natural speech production mechanism, i.e., the excitation source, the vocal tract system (i.e., filter) and the Source-Filter (S-F) interaction. The thesis aims to propose countermeasures that help in decreasing the % EER of the classification system both for the known attacks and for the unknown attacks on the ASV spoof 2015 challenge database. Furthermore, we explore the robustness of the features in various ways by evaluating the spoof dependency of the features and the performance of the features in the case of channel mismatch conditions using the Blizzard Challenge database.

7.2 Discussions

7.2.1 Performance of the Features on ASV Spoof Database

The ASV spoof challenge database consists of clean recording conditions with the similar set of recording conditions both in the training and in the testing phase. Considering the case of system-based features, these performed very well on the ASV spoof data. The MFCC could perform well for the known attacks and the cochlear filter-based (i.e., CFCCIF and CFCCIFS) features and the SBAE feature set detected unknown vocoder-independent spoof very well. Hence, their score-level fusion further reduced the EER of the SSD system. However, as known that the system-level information is not the only information in the speech signal, it is essential to study the role of source-based features as well.

In speech synthesis and voice conversion techniques, generally the spectral modification or spectral conversion it considered important during the speech generation process. The modeling of source characteristics has complementary information in enhancing the quality of machine-generated speech. Therefore, when features based on the strength of closure of the glottis, (i.e., *SoE*) and the nonlinear

prediction of the speech are used (which are rarely used in SS or VCS generation), a relative improvement in the performance of the SSD system is observed. It may be possible that if the source aspects are modeled in the machine-generated speech, then these features may not be able to capture the spoof-specific characteristics. The Fujisaki model is used extensively for prosody modeling in speech synthesis and hence, the features derived from the model may fail at times. An efficient way is to direct research towards the design of such features that are highly difficult to model in the synthetically generated speech. The use of the nonlinear S-F interaction is a possible attempt in this direction. There exist very few studies that explore the nonlinear coupling between the source and the system in a mathematical or a feature representative form. Hence, these features are found to be highly useful in detecting unknown vocoder-based spoofs even with lesser feature dimension.

7.2.2 Spoof Dependency of the Proposed Features

The general approach in evaluating the features or countermeasures is to use the entire training set of the ASV spoof challenge database which includes both SS and VCS spoofs. However, it may also be worthwhile to know if the training can be minimized as much as possible and still achieve reduced EER for any unknown attack. Given this, we attempt to train on individual spoofing algorithms and estimate the EER for the spoofs in the development and the evaluation set.

For the system-based features, on training with the VCS spoof only, both VCS and SS could be detected very well. With the use of SBAE features and training with the *S5* spoof, all the vocoder-based spoof conditions were detected with almost *~0 %* EER. On the other hand, the source-based features were not as robust as the system-based features and the source-based feature sets could identify only itself and the same type of attacks to some extent. This was mostly applicable to SS because its same type was generated using the same algorithm and the only change was in the number of utterances that were used in training. However, this is not the real case scenario. In the case of source-based features (such as F_0, *SoE1* and *SoE2*), the % EER decreased consistently with the increase in dynamic information for the different type of spoof. However, this decrease was not significant. This was even observed with the S-F interaction features. Thus, the system-based features could be a better countermeasure when only a single VCS spoof is available for training.

7.2.3 Robustness of the Features to Channel Mismatch Case

To evaluate the performance of the countermeasures for the channel mismatch case is the next immediate task that needs to be looked into even before robustness to signal degradation conditions is considered. This is because it is highly unlikely that the test spoofed speech will be of the same recording condition as that used in the training process. The evaluations on the Blizzard datasets show the vulnerability of the countermeasures to the channel mismatch case. Especially the system-based features that outperformed all available features on the ASV spoof challenge database did not generalize well for the channel mismatch case. Few vocoder-based speech recordings were detected with *~0* % EER while some had as high as *60-90* % EER. The same applied to unit-selected based spoofed speeches. Amongst the source-based features, even the prediction-based and the Fujisaki model-based features faced similar problems. The only consistency was observed in the F_0, *SoE1* and *SoE2* features with their dynamics that showed decreases in % EER with the increase in dynamic information. The relative decrease in % EER was more in statistical-based synthesis techniques as compared to unit-selection-based techniques. Hence, these showed significant robustness to the channel mismatch case. The use of residual energy features to capture the S-F interaction also showed less % EER, i.e., robustness to channel mismatch case. In the case of HMM-based speech synthesis, the % EER was less compared to the unit-selection, diphone and the hybrid approaches of speech generation. Hence, research can be further directed to develop countermeasures that handle the channel variability to a larger extent.

7.2.4 Performance of Humans *vs.* SSD systems

In [**123**], a study has been carried out that benchmarks automatic systems against human performance on speaker verification and spoofing detection tasks. The study was attempted to know whether human perceptual ability is important in identifying spoofing and whether humans can achieve better performance than SSD systems. A set of listening tests were designed to conduct speaker verification tasks and spoofing detection task. For the verification task, the ASV systems were found to outperform the humans. For the detection task, human listeners detect spoofing less well than most of the automatic approaches. However, humans are much better than any of the automatic countermeasures in detecting unit-selection based speech. It is

observed that the proposed SSD systems detected vocoder-based spoofs very well. However, human listeners generally fail to recognize the vocoder-based speech upon hearing. On the other hand, for the unit-selection speech, the SSD systems have shown high % EER. This is because this type of speech is similar to the natural speech except at the point of concatenation. However, humans can very easily perceive this discontinuity while listening and can identify the spoofed speech. Thus the SSD systems slightly contradict the human perception. In [**123**], it has been quantified that for vocoder-based speech, the performance of SSD systems is around four times better than humans to falsely accept spoofed speech as natural. While in the case of the vocoder-independent speech, the humans perform ten times better than the SSD systems. Thus, incorporating the use of perception mechanism to the feature extraction process is highly recommended. This is one of the reasons, due to which the proposed system-based features which use subband processing performs better in detecting the *S10* spoof.

7.3 Future Applications of the SSD task

Several applications of the SSD task have been mentioned in Section 1.5 amongst which the immediate applicable is the security of ASV systems for reliable telephone banking, personal identification and computer logins, etc. This includes the design of countermeasures and using it jointly with the ASV systems. The performance of the ASV systems should not be affected when the spoof detection system is incorporated into the ASV framework. Although significant progress is made for detecting spoofed speech in clean environment, the problem is yet to be solved especially for signal degradation and channel mismatch conditions.

Amongst the several other applications, an important area would be to use the features for its counter application. That is to improve the quality of the synthetic and voice converted speech by using the knowledge of the lacking features in the spoofed speech. An example of which is the phase of the signal. The vocoder-based speech is known to lack the phase information as only the spectral magnitude is processed for speech generation. Knowing this and incorporating the phase-based information in some form during the speech generation process can aid to make the speech sound more natural. However, this is again a threat to the spoof detector systems which further needs to be modified and generalized.

Another important application of the design of countermeasures is the possible development of objective measures for the evaluation of TTS and voice conversion systems. In the evaluation of the TTS and voice conversion system, the subjective tests play an important role. Although the end users of the TTS or voice conversion system applications are humans, they may not be always efficient in evaluating these systems. Humans are only able to judge if the machine-generated speech is natural or not and that too in a naïve way (especially, if they are not speech processing experts). That is, they cannot identify the cause of degraded voice quality which may be due to lack of spectral information, prosody information, improper pitch modeling and many other acoustic properties which are not modeled properly, etc. Hence, using the countermeasures could be highly essential for objective evaluation and could aid in avoiding to some extent, if not eliminating the extensive subjective evaluations that are carried out to test the system performance.

7.4 Contributions from the Thesis

- Instead of considering the traditional approach for the known and unknown attacks, we bring the approach of the known, same and different type of spoofing attacks. That is, how well training with synthetic speech can test the voice conversion spoofs and vice-a-versa. This can be an efficient way of evaluating the countermeasures.
- The proposal of system-based features, i.e., both the CFCCIF and SBAE are found to reduce significantly the % EER of vocoder-independent spoofs like MARY TTS (i.e., the *S10* spoof of the ASV spoof database).
- A simple yet a novel approach to explore the dynamics of F_0 and *SoEs* was presented and this approach was found to be very effective for identifying vocoder-based spoofs. This approach proved to be a generalized feature set even for the channel mismatch case.
- Considering the nonlinearities in the samples of speech, the nonlinear prediction of speech was explored. The nonlinearity in the spoofed speech is almost equal to the linearity in a spoofed speech which is not the case in natural speech. Using this finding, a combination of NLP with LP or LTP gave better results along with the traditional LP-LTP combination.

- The Fujisaki model was explored for the first time to identify the lacking prosodic information in the speech as compared to its traditional use in prosody modification for TTS systems.
- An implicit way of using nonlinear coupling present in the speech production mechanism was proposed. This approach was found to detect vocoder-based spoofs even with less feature dimension. In addition, it was observed that the residue information was resilient enough to detect the spoofed speech in the case of signal degradation conditions.

7.5 Limitations of the Present Research Work

- Only system-based features worked to certain extend on *S10* USS spoof. Such low EER was not observed with source-based features or source-system interaction features.
- The system-based features were robust to unknown spoofs, however, not robust enough to channel mismatch conditions.
- Fujisaki model based features may not work well if the spoofed speech has used prosody modeling during the speech generation process. In addition, the generalization to non-parallel utterances is even more difficult.
- The channel mismatch was attempted; however, no ground truth regarding the actual channel condition is available.
- The analysis of the thesis is more specific to the ASV spoof and the claims do not hold directly for the Blizzard data.

7.6 Research Issues and Future Research Directions

Based on the literature presented about spoofing attacks (with and without ASV systems for known and unknown attacks), various research issues or gap areas in the SSD task can be bought out in the following aspects.

Diversity of spoofing attacks: The ASV spoof database consists of a large number of spoofing attacks. However, majority of the spoofing algorithm include variations of the VC techniques and less of the SS techniques. Recently, there have been significant development in using DNNs for SS techniques [**109**], [**110**] that could also be included as spoofing techniques. The use systems of Blizzard dataset allows to use

recent techniques in SS. However, these are available for very few speakers and utterances. For VCS such an evaluation is currently not carried out. The use of VC systems of the recently organized Voice Conversion challenge [**234**] can be an initial step towards evaluation of the features to unknown VC spoofs. Recently, in [**235**], the authors show that noisy yet intelligible speech can degrade the performance of ASV systems. This can be explored as a potential spoof to be detected.

Direct and physical access: The ASV spoof database does not consider the case of physical access. The physical access is the actual spoofing where the speech is played back through a microphone into the ASV system. The recently developed Audio-Visual (AV) spoofing database includes ten realistic spoofing threats generated using replay, SS and VC [**236**]. Although this database consists of *44* speakers that is less as compared to the ASV spoof database, the AV spoof database makes provision of providing the physical access spoofing attacks to evaluate the countermeasures.

Number of speakers: The ASV spoof database consists of a large number for male and female speakers for evaluation as compared to the AV spoof database. The number of speakers available through Blizzard challenge dataset is further less. The effect of speaker number can be both while training and testing stage. In [**70**], % EER showed improvements as the number of speakers used for training increases. However, this changes with the type of spoofing attacks and features. Thus, the anti-spoofing measures must also justify their independence or robustness to the number of speakers and the speaker's voice under consideration.

Generalization of features: Generally, vocoder-based spoofs are studied and phase-based features were initially used. However, with the development of phase-aware vocoders (such as AHOCODER [**108**]), the phase-based features may not be effective. Thus, features other than phase-based approaches need more attention. The use of features based on subband processing which showed better performance due to embedded perceptual information can be explored. The basic idea is that the features should be able to detect any spoofed speech generated from any algorithm, irrespective of the availability of the spoof in the training set.

Channel mismatch conditions: It is highly unlikely that the test spoofed speech will be of the same recording condition as used in the training process. The evaluations on the Blizzard datasets show the vulnerability of the countermeasures to the channel mismatch case. Again, the use of Blizzard dataset is not sufficient due to less number of utterances and less number of speakers and the lack of VC systems for evaluation. The use of AV spoof database having session variability and which has significant speakers can also be explored.

Signal degradation conditions: The available ASV and AV spoof databases considers spoofing under clean conditions only. A noisy database is developed by adding background noises at various SNR to the ASV spoof challenge 2015 database in [**95**]. This database can be generated for the AV spoof database considering the physical access case as well. It needs to be studied further how the diversity in the various noise types affects the performance of the SSD system. Rather than evaluating the existing features for noisy dataset, the countermeasures must be modified to be robust to signal degradation as well.

Performance with ASV systems: Studies have reported alteration in the performance of the baseline ASV systems when used with the countermeasures. However, the performance of the ASV system should not depend on its joint use of countermeasures. In addition, generally, for spoof detection, the FAR is considered. However, the features should have lower FRR when used with ASV systems to provide better user convenience by lesser rejections of genuine trials.

Performance evaluation: The current evaluation system estimates the EER for each of the spoofing algorithm and then averages the EERs (attack-dependent EER). However, the realistic approach is to use the attack-independent EER that assumes only two classes of natural and spoofed speech. Based on this EER, the individual EERs can be obtained by identifying if the likelihood scores for a particular spoof (from a specific spoofing algorithm) was greater than or less the threshold. Such an approach allows to have a fixed threshold which is a more realistic scenario in case of unknown attacks. Hence, research can be directed towards using the attack-ndependent approach to estimate the EER.

References

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

[2] A. A. Paulino, "Contributions to biometric recognition: Matching identical twins and latent fingerprints," Ph.D. Thesis, Michigan State University, MI, USA, 2013.

[3] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 475-487, April 1976.

[4] International Organization for Standardization (ISO). ISO/IEC 30107-1:2016(en): Information technology-Biometric presentation attack detection-Part 1: Framework. [Available Online]. https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-1:ed-1:v1:en {Last accessed:26 Aug. 2016}.

[5] D. A. Reynolds, "Automatic Speaker Recognition using Gaussian Mixture Speaker Models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173-192, 1995.

[6] S. Furui, "An Overview of Speaker Recognition Technology," in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Springer US, 1996, ch. 2, pp. 31-56.

[7] A. Neustein and H. A. Patil, *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. Springer, Oct. 2011.

[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acous., Speech, and Sig. Process.*, vol. 28, no. 4, pp. 357-366, 1980.

[9] H. A. Patil, "Speaker Recognition in Indian Languages: A Feature Based Approach," Ph.D. Thesis, Dept. of Elec. Engg., IIT Kharagpur, 2005.

[10] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 16, no. 5, pp. 980-988, 2008.

[11] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010: The Speaker and Lang. Recog. Workshop*, Brno, Czech Republic, 2010, pp. 14-24.

[12] H. A. Patil and T. K. Basu, "LP spectra vs. Mel spectra for identification of professional mimics in Indian languages," *Int. Jour. Speech Tech. (IJST), Springer-Verlag*, vol. 11, no. 1, pp. 1-16, Mar. 2008.

[13] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Odyssey 2006: The Speaker and Lang. Recognition Workshop*, San Juan, 2006, pp. 1-6.

[14] D. Gomathi, S. A. Thati, K. V. Sridaran, and B. Yegnanarayana, "Analysis of mimicry speech," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, Oregon, 2012, pp. 695-698.

[15] M. M. Chakka, et al., "Competition on counter measures to 2-D facial spoofing attacks," in *IEEE Int. Joint Conf. on Biometrics (IJCB)*, Washington DC, Oct. 2011, pp. 1-6.

[16] G. Marcialis, et al., "First international fingerprint liveness detection competition - LivDet 2009," in *Image Analysis and Processing ICIAP*, P. Foggia, C. Sansone, and M. Vento, Eds. Berlin Heidelberg: Lect. Notes in Comp. Sc. (LNCS)-Springer, 2009, pp. 12-23.

[17] D. Yambay, J. S. Doyle, K. W. Bowyer, A. Czajka, and S. Schuckers, "LivDet-Iris 2013 - Iris liveness detection competition 2013," in *IEEE Int. Joint Conf. on Biometrics (IJCB)*, Florida, USA, 2014, pp. 1-8.

[18] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 925-928.

[19] Z. Wu, et al., "SAS: A speaker verification spoofing database containing diverse attacks," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 4440-4444.

[20] Z. Wu, et al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2037-2041.

[21] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, 2nd ed. Pearson, 2012.

[22] ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge, Theme in the 2017 edition: Audio replay attack detection. [Available Online]. http://www.spoofingchallenge.org/ {Last accessed: 26 April 2017}.

[23] K. S. Rao, *Predicting Prosody from Text for Text-to-Speech Synthesis*, 1st ed. New York, 2012.

[24] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, Oregon, USA, 2012, pp. 1-11.

[25] K. Prahallad, et al., "The Blizzard Challenge 2014," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Singapore, 2014, pp. 1-14.

[26] Z. Wu, et al., "Spoofing and countermeausres for speaker verification: A survey," *Speech Comm.*, vol. 66, pp. 130-153, 2015.

[27] A. Paul, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *Int. Conf. on Sig. Process. and Comm. (SPCOM)*, Bangalore, India, 2016, pp. 1-5.

[28] J. A. Villalba, "Advances on speaker Recognition in non collaborative environments," Ph.D. Thesis, University of Zaragoza, Spain, 2014.

[29] D. H. Klatt, "Review of text-to-speech conversion for English," *Jour. of Acoust. Soc. of Amer. (JASA)*, vol. 82, no. 3, pp. 737-793, 1987.

[30] K. A. Lenzo and A. W. Black, "Diphone collection and synthesis," in *Int. Conf. of Spoken Lang. Process. (ICSLP)*, Beijing China, 2000, pp. 1-4.

[31] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, vol. 11, Tokyo, 1986, pp. 2015-2018.

[32] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Atlanta, Georgia, 1996, pp. 373-376.

[33] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *IEICE Trans. Information and Systems*, vol. J83-D-II, pp. 2099-2107, 2000.

[34] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Int. Conf. on Spoken Lang. Process. (ICSLP)*, Pittsburgh, Pennsylvania, 2006, pp. 1762-1765.

[35] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039-1064, Nov. 2009.

[36] J. Yamagishi, T. Kobayashi, Y. Nakona, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, no. 1, pp. 66-83, Jan.2009.

[37] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," in *4th ISCA Workshop on Speech Synthesis*, Blair Atholl, Scotland, 2001, pp. 1-6.

[38] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice Conversion: Factors responsible for quality," in *IEEE Int. Conf. Acous., Speech, and Sig. Process. (ICASSP)*, Tampa, Florida, USA, 2007, pp. 513-516.

[39] K. Shikano, K. Lee, and R. Reddy, "Speaker adaptation through vector quantization," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Tokyo, Japan, 1986, pp. 2643-2646.

[40] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Seattle, Washington, USA, 1988, pp. 655-658.

[41] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131-142, Mar. 1998.

[42] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Seattle, Washington, USA, 1988, pp. 285-288.

[43] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Salt Lake City, UT, 2001, pp. 841-844.

[44] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 18, no. 5, pp. 922-931, 2010.

[45] D. Sündermann, et al., "Text-independent voice conversion based on unit selection," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process., (ICASSP)*, Toulouse, France, 2006, pp. 81-84.

[46] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 3057-3061.

[47] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 18, no. 5, pp. 954-964, Apr. 2010.

[48] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted boltzmann machines," *IEEE/ACM. Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 3, pp. 580-587, 2015.

[49] G. Ben and S. King, "Transforming F0 contours," in *Euro. Conf. on Speech Comm. and Tech. (EUROSPEECH)*, Geneva, Switzerland, 2003, pp. 101-104.

[50] E. E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Honululu, Hawaii, USA, 2007, pp. 509-512.

[51] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 14, no. 4, pp. 1109-1116, 2006.

[52] D. Lolive, N. Barbot, and O. Boeffard, "Pitch and duration transformation with non-parallel data.," in *Speech Prosody*, Campinas, Brazil, 2008, pp. 111-114.

[53] D. Genoud and G. Chollet, "Speech pre-processing against intentional imposture in speaker recognition," in *Int. Symposium on Chinese Spoken Lang. Process. (ISCSLP)*, Sydney, Australia, 1998, pp. 1-11.

[54] B. L. Pellom and J. H. L. Hansen, "An experimental study of speaker verification sensitivity to compute voice-altered imposters," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Phoenix, Arizona, 1999, pp. 837-840.

[55] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech Synthesis using HMMs with dynamic features," in *IEEE Int. Conf. on Acous. Speech, and Sig. Process. (ICASSP)*, Atlanta, Georgia, 1996, pp. 389-392.

[56] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *ESCA/COCOSDA Int. Workshop on Speech Synthesis*, 1998, pp. 273-276.

[57] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Euro. Conf. Speech Comm. Tech. (EUROSPEECH)*, Budapest, Hungary, 1999, pp. 1223-1226.

[58] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *IEEE Int. Conf. on Acous. Speech, and Sig. Process. (ICASSP)*, Munich, Germany, 1997, pp. 1611-1614.

[59] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - A study of technical impostor techniques," in *Euro. Conf. Speech Comm. Tech. (EUROSPEECH)*, Budapest, Hungary, 1999, pp. 1211-1214.

[60] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis*, 2002, pp. 227-230.

[61] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Toulouse, May 2006, pp. 933-936.

[62] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Antwerp, Belgium, 2007, pp. 2053-2056.

[63] T. Kinnunen, et al., "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Kyoto, 2012, pp. 4401-4404.

[64] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouche, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788-798, 2011.

[65] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. on Fund. of Electronics, Comm. and Comp. Sciences*, vol. 88-A, pp. 280-286, 2005.

[66] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistcs derived from image analysis," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, USA, 2012, pp. 370-373.

[67] P. L. D. Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *IEEE Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 4844-4847.

[68] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, Oregon, USA, 2012, pp. 1700-1703.

[69] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *IEEE Int. Conf. on Acous. Speech, and Sig. Process. (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 7234-7238.

[70] J. Sanchez, et al., "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Info. Forensics and security*, vol. 10, no. 4, pp. 810-820, April 2015.

[71] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2082-2086.

[72] X. Xiao, et al., "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2052-2056.

[73] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2092-2096.

[74] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2042-2046.

[75] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Speech Comm.*, vol. 81, pp. 30-41, Jul. 2016.

[76] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2077-2081.

[77] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasure challenge 2015," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2072-2076.

[78] S. Weng, et al., "The SYSU system for the Interspeech 2015 automatic speaker verification spoofing and countermeasures challenge," arXiv:1507.06711, 2015.

[79] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection-The SJTU system for ASVspoof 2015 challenge," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2097-2101.

[80] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2064-2071.

[81] T. B. Patel and H. A. Patil, "Combining evidences from Mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2062-2066.

[82] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5475-5479.

[83] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparision of features for synthetic speech detection," in *Int. Speech Comm Assos. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2087-2091.

[84] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 2119-2123.

[85] C. Zhang, et al., "Joint information from nonlinear and linear features for spoofing detection: An i-vector/DNN based approach," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5035-5039.

[86] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F0) and strength of excitation (SoE) for spoofed speech detection," in *Int. Conf. on Acous. Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5105-5109.

[87] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Odyssey 2016: The Speaker and Lang. Recognition Workshop*, Bilbao, Spain, 2016, pp. 270-276.

[88] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-Spoofing: Constant Q cepstral coefficients," in *Odyssey 2016: The Speaker and Lang. Recognition Workshop*, Bilbao, Spain, 2016, pp. 283-290.

[89] M. H. Soni, T. B. Patel, and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 1820-1824.

[90] H. Bhavsar, T. Patel, and H. Patil, "Novel nonlinear prediction based features for spoofed speech detection," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 155-159.

[91] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *in IEEE Journal of Selected Topics in Signal Processing (JSTSP), Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification (accepted).*

[92] Y. Qian, N. Chen, and K. Yu, "Deep Features for Automatic Spoof Detection," *Speech Comm.* , vol. 85, pp. 43-52, 2016.

[93] T. B. Patel and H. A. Patil, "Significance of source-filter interaction for classification of natural vs. spoofed speech," *in IEEE Journal of Selected Topics in Signal Processing (JSTSP), Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification (accepted).*

[94] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection under noisy conditions: A preliminary investigation and an initial database," http://arxiv.org/pdf/1602.02950v1.pdf, 2016.

[95] X. Tian, Z. Wu, X. Xiao, and H. Li, "An investigation of spoofing speech detection under additive noise and reverberant conditions," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 1715-1719.

[96] C. Hanilci, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in presence of additive noise," *http://arxiv.org/abs/1603.03947v2*, pp. 1-21, May 2016.

[97] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Int. Speech Comm. Assoc.*, Scandinavia, 2001, pp. 759-761.

[98] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Salt Lake City, UT, 2001, pp. 805-808.

[99] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *Int. Symposium on Chinese Spoken Lang. Process. (ISCSLP)*, Taiwan, 2010, pp. 309-312.

[100] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 20, no. 8, pp. 2280-2290, Oct. 2012.

[101] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: The telephone speech case," in *Asia-Pacific Sig. & Infor. Process. Assoc. Annual Summit and Conf. (APSIPA ASC), 2012*, Hollywood, CA, Dec. 2012, pp. 1-5.

[102] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, OR, USA, 2012, pp. 1688-1691.

[103] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 940-944.

[104] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Arlington, VA, 2013, pp. 1-8.

[105] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and anti-spoofing in the i-vector space," *IEEE Trans. Info. Foren. and Sec.*, vol. 10, no. 4, pp. 821-832, February 2015.

[106] S. Ganpathy, "Factor analysis method for joint speaker verification and spoof detection," in *IEEE Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, New Orleans, USA, 2017, pp. 5385-5389.

[107] M. Sahidullah, H. Delgado, M. Todisco, and Z.-H. Tan, "Integrated spoofing countermeasures and automatic speaker verification: an evaluation on ASVspoof 2015," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 1700-1704.

[108] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Italy, 2011, p. 1809–1812.

[109] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Vancouver, Canada, 2013, pp. 7962-7966.

[110] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *IEEE Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5140-5144.

[111] A. Black, P. Taylor, and R. Caley. (1988) The Festival speech synthesis system. [Available Online]. http://festvox.org/festival/ {Last accesssed: 24 Aug. 2014}.

[112] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Comm.*, vol. 36, no. 1, pp. 31-43, Jan. 2002.

[113] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Euro. Conf. Speech Process. Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 601-604.

[114] H. A. Patil, "TTS Lecture Module 1-5," in UGC's e-PG Pathshala Programme for e-Content Creation, 2015.

[115] P. Taylor, *Text-to-Speech Synthesis*, 1st ed. Cambridge University Press, 2009.

[116] J. Yamagishi. (2006) An Introduction to HMM-Based Speech Synthesis. [Available Online]. https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/TrajectoryModelling/HTS-Introduction.pdf {Last accessed: 19 Sept. 2016}.

[117] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayshi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based Speech Synthesis System," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Istanbul, Turkey, 2000, pp. 1315-1318.

[118] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation Filter for Speech Synthesis," *Trans. IECEJ (in Japanese)*, vol. J66, no. A, pp. 10-18, 1983.

[119] Y. Stylianou, "Voice Transformation: A survey," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 3585-3588.

[120] A. F. Machado and M. Queiroz, "Voice Conversion: A critical Survey," in *Proceedings of Sound and Music Computing (SMC)*, 2010, pp. 1-8.

[121] International Telecommunication Union: A Method for subjective performance Assessment of the quality of speech voice output devices, ITU-T Rec. P.85. [Available Online]. https://www.itu.int/rec/T-REC-P.85/en {Last accessed: 25 Apr. 2017}.

[122] C. Benoît, M. Griceb, and V. Hazanc, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, no. 4, pp. 381-392, Jun. 1996.

[123] Z. Wu, et al., "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 24, no. 4, pp. 768-783, 2016.

[124] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *IEEE Pacific Rim Conf. on Comm., Comp. and Sig. Process.*, Victoria, BC, 1993, pp. 125-128.

[125] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acous., Speech and Sig. Process.*, vol. 26, no. 1, pp. 43-49, Feb. 1978.

[126] F. Hinterleitner, S. Möller, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from Blizzard Challenges 2008 and 2009," in *Blizzard Challenge*, Kansai Science City, Japan, 2010, pp. 1325-1328.

[127] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Moller, "Instrumental assessment of prosodic quality for text-to-speech signals," *IEEE Signal Process. Letters*, vol. 19, no. 5, pp. 255-258, May 2012.

[128] H. Patil, et al., "Algorithm for speech segmentation at syllable level for text-to-speech synthesis system in Gujarati," in *Oriental Int. Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) Conf.*, Gurgaon, India, 2013, pp. 1-7.

[129] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 3, pp. 232-239, Mar. 2001.

[130] A. W. Black, H. Zen, and K. Tokuda, "Statistical Parametric Speech Synthesis," in *Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Honolulu, Hawaii, USA, 2007, pp. 1229-1232.

[131] J. Nurminen, H. Silén, V. Popa, E. Helander, and M. Gabbouj , "Voice Conversion," in *Speech Enhancement, Modeling and Recognition - Algorithms and Applications*, S. Ramakrishnan, Ed. InTech, 2014, ch. 5, pp. 69-94.

[132] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, no. 3-4, pp. 187-207, Apr. 1999.

[133] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, San Francisco, CA, 1992, pp. 137-140.

[134] Voice Transformation [Available Online]. http://www.festvox.org/ {Last accessed 19th Sept. 2016}.

[135] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximumlikelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 8, pp. 2222-2235, 2007.

[136] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice econversion based on tensor representation of speaker space," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Florence, Italy, 2011, pp. 653-656.

[137] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least square regression," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 20, no. 3, pp. 806-815, Mar. 2012.

[138] The MARY Text-to-Speech System (MaryTTS). [Available Online]. http://mary.dfki.de/ {Last accessed: 24 Aug. 2015}.

[139] Speech Synthesis Special Interest Group (SynSIG): Blizzard Challenge. [Available Online]. http://www.synsig.org/index.php/Blizzard_Challenge {Last accessed: 3 Sept. 2016}.

[140] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006(1-12), 2014.

[141] Submissions and listening test results from previous Blizzard Challenges. [Available Online]. http://www.cstr.ed.ac.uk/projects/blizzard/data.html {Last accessed: 12 Aug. 2016}.

[142] S. P. Kishore and A. W. Black, "Unit Size in Unit Selection Speech Synthesis," in *Euro. Conf. Speech Process. Technology (EUROSPEECH)*, GENEVA, 2003, pp. 1317-1320.

[143] S. Talesara, "Design of syllable-based speech segmentation methods for text-to-speech (TTS) system in Gujarati," M.Tech. Thesis, DA-IICT, Gandhinagar, 2013.

[144] P. G. Deivapalan, M. Jha, R. Guttikonda, and H. A. Murthy, "DONLabel: An Automatic Labeling Tool for Indian Languages," in *The Nat. Conf. on Comm. (NCC)*, IIT Bombay, 2008, pp. 263-266.

[145] S. Talesara, H. Patil, T. Patel, H. Sailor, and N. Shah, "A novel gaussian filter-based automatic labeling of speech data for TTS system in Gujarati language," in *Int. Conf. on Asian Lang. Process. (IALP)*, Urumki, China, August 17-19, 2013, pp. 139-142.

[146] H. Patil, et al., "A syllable-based framework for unit selection synthesis in 13 Indian languages," in *Oriental Int. Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) Conf.*, Gurgaon, INDIA, 2013, pp. 1-8.

[147] Nagoya Institute of Technology. [Available Online]. http://hts.sp.nitech.ac.jp/ {Last accessed: 24 Aug. 2014}.

[148] N. J. Shah, "HMM-based speech synthesis system for Gujarati language," M.Tech. Thesis, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India, July, 2013.

[149] B. Ramani, et al., "A common attribute based unified HTS framework for speech synthesis in Inidan languages," in *ISCA workshop on Speech Synthesis*, Barcelona, Spain, pp. 311-316, 2013.

[150] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, "Effectiveness of PLP-based phonetic segmentation for speech synthesis," in *IEEE Int. Conf. on Acous. Speech, and Sig. Process. (ICASSP)*, Florence, 2014, pp. 270-274.

[151] S. King, "An introduction to statistical parametric speech synthesis," *Sadhna, Indian Academy of Sciences*, vol. 36, no. 5, pp. 837-852, Oct. 2011.

[152] H. A. Patil, M. C. Madhavi, K. D. Malde, and B. B. Vachhani, "Phonetic transcription for fricatives and plosives for Gujarati and Marathi languages," in *Int. Conf. on Asian Lang. Process. (IALP)*, Hanoi, Vietnam, 2012, pp. 177-180.

[153] N. Shah, H. Patil, M. Madhavi, H. Sailor, and T. Patel, "Deterministic annealing EM algorithm for developing Gujarati TTS system," in *Int. Symposium on Chinese Spoken Lang. Process. (ISCSLP)*, Singapore, 2016, pp. 526-530.

[154] D. Reynolds, "Gaussian Mixture Models," *Encyclopedia of Biometric Recognition, Springer, Journal Article*, pp. 1-5, Feb. 2008.

[155] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Jour. of the Royal Stat. Soc. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[156] DET-Curve Plotting software for use with MATLAB. [Available Online].
http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm {Last accessed: 13 April 2015}.

[157] A. Martin, G. Doddington, T. Kamm, and M. Ordowski, "The DET curve in assessment of detection task performance," in *Euro. Conf. Speech Process. Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 1895-1898.

[158] Figure for Anatomy of Ear. [Available Online]. http://www.hearingpro.com.au/anatomy-of-the-human-ear.html {Last accessed: 13 April 2015}.

[159] The Cochlea in the human ear. [Available Online].
https://introtohearingscience.wordpress.com/ {Last accessed: 13 April 2015}.

[160] J. H. L. Hansen and D. T. Chappell, "An auditory-based distortion measure with application to concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 489-495, 1998.

[161] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Comm.*, vol. 45, no. 2, pp. 153-170, Feb. 2005.

[162] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency," *Speech Comm.*, vol. 53, no. 5, pp. 707-715, May 2011.

[163] S. Shamma and D. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 107, no. 5, pp. 2631-2644, May 2000.

[164] D. O'Shaughnessy, *Speech communication: Human and machine*, 1st ed. Addison-Wesley, 1987.

[165] Q. Li, "An auditory-based transform for audio signal processing," in *IEEE Workshop on Applications of Sign. Process. to Audio and Acous.*, New Paltz, NY, 2009.

[166] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 6, pp. 1791-1801, 2011.

[167] N. Singh, N. Bhendwade, and H. A. Patil, "Novel cochlear filter based cepstral coefficients for classification of unvoiced fricatives," *Int. Jour. on Natural Lang. Computing*, vol. 3, no. 4, pp. 21-40, Aug. 2014.

[168] S. G. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1998.

[169] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 118, no. 2, pp. 887-906, Aug. 2005.

[170] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, vol. 416, no. 6876, pp. 87-90, 2002.

[171] S. Furui, "On the role of spectral transistion for speech perception," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 80, no. 4, pp. 1016-1025, 1986.

[172] Q. Lin, E.-E. Jan, C. Che, D.-S. Yuk, and J. Flanagan, "Selective use of speech spectrum and VQGMM method for speaker identification," in *Euro. Conf. Speech Process. Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 2415-2418.

[173] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[174] N. Jaitly and G. E. Hinton, "A new way to learn acoustic events," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1-9, 2011.

[175] N. Jaitly and G. E. Hinton, "Using an autoencoder with deformable templates to discover features for automated speech recognition," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 1737-1740.

[176] S. Takaki and J. Yamagishi, "A deep auto-encoder based lowdimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5090-5094.

[177] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. Princeton University Press, 1987.

[178] I. R. Titze, *Principles of Voice Production*. Prentice Hall (currently published by NCVS.org), 1994.

[179] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Sig. Process. Letters*, vol. 16, no. 9, pp. 469-472, 2009.

[180] P. Bachhav, H. A. Patil, and T. B. Patel, "A novel filtering based approach for epoch extraction," in *IEEE Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 4784-4788.

[181] D. Gandhi, T. B. Patel, and H. A. Patil, "A novel lowpass filtering-based approach for estimating strength of excitation from speech signal," in *Int. Conf. on Sig. Process. and Comm. (SPCOM)*, Bangalore, 2016.

[182] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Speech and Audio Process.*, vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[183] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse," *Speech Comm.*, vol. 11, no. 2-3, p. 109–118, 1992.

[184] T. Raitio, et al., "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 153-165, 2011.

[185] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech wave," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 50, no. 2B, pp. 637-655, 1971.

[186] L. Ljung, *System Identification-Theory for the User*, 2nd ed. Prentice Hall, 1998.

[187] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154-161, 2006.

[188] J. Markhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561-580, Apr. 1975.

[189] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. of Acous., Speech and Sig. Process.*, vol. 27, no. 4, pp. 309-319, Aug. 1979.

[190] D. D. Mehta and P. J. Wolfe, "Statistical properties of linear prediction analysis underlying the challenge of formant bandwidth estimation," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 137, no. 2, p. 944–950, Feb. 2015.

[191] M. R. Schroder and B. S. Atal, "Code-excited Linear Prediction (CELP): High-quality speech at low rates," in *IEEE Int. Conf. on Acous. Speech, and Sig. Process. (ICASSP)*, vol. 10, Tampa, Florida, USA, 1985, pp. 937-940.

[192] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri, "Temporal masking for bit-rate reduction in audio codec based on frequency domain linear prediction," in *IEEE Int. Conf. on Acous., Speech and Sig. Process., (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 4781-4784.

[193] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd ed. Wiley & Sons, 2008.

[194] "Digital cellular telecommunications systems (Phase 2+) (GSM); Full rate speech," Transcoding GSM 06.10 ETSI Std., Rev. version 8.1.1, 1999.

[195] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*, 2nd ed. John Wiley & sons, 2000.

[196] M. J. Korenberg, "Indentifying nonlinear difference equation and functional expansion representation: The fast orthogonal algorithm," *Annals of Biomedical Engineering*, vol. 16, pp. 123-142, 1988.

[197] H. A. Patil and T. B. Patel, "Nonlinear Prediction of Speech by Volterra-Wiener Series," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 1687-1691.

[198] M. Chetouani, A. Hussain, B. Gas, M. Milgram, and J.-L. Zarader (Eds.), "Advances in Nonlinear Speech Processing," in *Int. Conf. on Nonlinear Speech Process (NOLISP), Lecture Notes in Artif icial Intelligence (LNAI), Springer*, 2007.

[199] H. Patil and T. Patel, "Chaotic mixed excitation source for speech synthesis," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, Singapore, 2014, pp. 785-789.

[200] G. H. Alipoor and M. H. Savoji, "Speech coding using non-linear prediction based on Volterra series expansion," in *13th Int. Conf. on Speech and Computer, SPECOM*, 2006, pp. 367-370.

[201] H. Fujisaki, "Information, prosody, and modeling," in *Proceedings of Speech Prosody*, Nara, Japan, March 2004, pp. 1-10.

[202] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations," Dept. for Speech, Music and Hearing , KTH Stockholm, Quarterly Progress and Status Report, 1981.

[203] H. Fujisaki, S. Ohno, and W. Gu, "Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command–response model for generation of their F0 contours," in *Int. Sym. on Tonal Aspects of Lang.—with Emphasis on Tone Lang.*, Beijing, China, 2004, pp. 61-64.

[204] A. Rajpal, et al., "Native language identification using spectral and source-based features," in *Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 2383-2387.

[205] A. Sakurai and H. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," in *4th Int. Conf. on Spoken Lang. Process., (ICSLP)*, vol. 2, Philadelphia, PA, 1996, pp. 817-820.

[206] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process., (ICASSP)*, 2002, pp. 509-512.

[207] T. B. Patel and H. A. Patil, "Analysis of natural and synthetic speech using Fujisaki model," in *IEEE Int. Conf. Acous., Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5250-5254.

[208] A. E. Aronson and D. M. Bless, *Clinical Voice Disorders*. New York: Thieme Medical Publishers, 2009.

[209] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 133, no. 15, pp. 3050-3061, May 2013.

[210] T. Patel and H. Patil, "Novel approach for estimating length of the vocal folds using Fujisaki model," in *Int. Symposium on Chinese Spoken Lang. Process. (ISCSLP)*, Singapore, 2012, pp. 308-312.

[211] CMU-ARCTIC Speech Synthesis Database. [Available Online].
http://festvox.org/cmu_arctic/index.html {Last accessed: 20 March, 2013}.

[212] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice-Hall, Eaglewood Cliffs, 1988.

[213] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Labs Technical Journal (BLTJ)*, vol. 52, pp. 1233-1268, Jul. 1972.

[214] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *KTH Quaterly Progress and Status Report (STL-QPSR)*, vol. 26, no. 4, pp. 001-013, 1985.

[215] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *KTH Quaterly Progress and Status Report (STL-QPSR)*, vol. 36, no. 2-3, pp. 119-156, 1995.

[216] J. P. S. R. Cabral, "HMM-based speech synthesis using an acoustic glottal source model," Ph.D. Thesis, The Centre for Speech Technology Research, School of Informatics, University of Edinburgh, 2010.

[217] T. Yoshimura, K. Tokuda, T. Masukom, Kobayashi, T, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Euro. Conf. Speech Comm. Tech. (EUROSPEECH)*, Aalborg, Denmark, 2001, pp. 2263-2266.

[218] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and groupdelay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Workshops on Models and Analysis of Vocal Emissions for Bio. Appl. (MAVEBA)*, Firenze, Italy, 2001, pp. 1-6.

[219] Y. Stylianou, "Concatenative speech synthesis using a harmonic plus noise model," in *3rd ESCA Speech Synthesis Workshop (SSW)*, Jenolan Caves, 1998, pp. 261-266.

[220] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in *Int. Conf. on Spoken Lang. Process. (ICSLP)*, Yokohama, Japan, 1994, pp. 1043-1046.

[221] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Int. Conf. on Acou. Speech and Sig. Process. (ICASSP)*, Pennsylvania, USA, 2005, pp. 9-12.

[222] L. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Comm.*, vol. 28, no. 3, pp. 211-226, 1999.

[223] M. D. Plumpe, "Modeling of the glottal flow derivative waveform with application to speaker identification," Masters Thesis, MIT, Dept. of Electrical Engg. and Commputer Science, Feb. 1997.

[224] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. on Speech and Audio Process.*, vol. 7, no. 5, pp. 569-586, Sep. 1999.

[225] I. R. Titze and A. Palaparthi, "Sensitivity of Source–Filter Interaction to Specific Vocal Tract Shapes," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2507-2515, Dec. 2016.

[226] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *Jour. Acoust. Soc Amer. (JASA)*, vol. 123, no. 5, pp. 2733-2749, May 2008.

[227] I. Titze, T. Riede, and P. Popolo, "Nonlinear source-filter coupling in phonation: Vocal excercises," *Jour. Acoustic Soc. Amer. (JASA)*, vol. 123, no. 4, pp. 1902-1915, Apr. 2008.

[228] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Comm.*, vol. 1, no. 3-4, pp. 167-184, Dec. 1982.

[229] C. R. Jankowski, "Fine structure features for speaker identification," Ph.D. Thesis, MIT, Lexington, MA, USA, 1996.

[230] J. Kane, I. Yanushevskaya, A. N. Chasaide, and C. Gobl, "Exploiting time and frequency domain measures for precise voice source parameterisation," in *Proc. Speech Prosody*, Shanghai, China, 2012, pp. 143-146.

[231] J. Kane. Voice_Analysis_Toolkit: A set of MATLAB codes for carrying out glottal source and voice quality analysis. [Available Online]. https://github.com/jckane/Voice_Analysis_Toolkit {Last accessed: 06 Dec. 2016}.

[232] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Brisbane, Australia, 2014, pp. 960-964.

[233] D. Pati and S. R. M. Prasanna, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," *Int J Speech Technol (IJST)*, vol. 18, pp. 333-350, 2015.

[234] T. Toda, et al., "The voice conversion challenge 2016," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, San Fransisco, USA, 2016, pp. 1632-1636.

[235] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *IEEE Int. Conf. on Acous. Speech, and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5500-5505.

[236] S. K. Ergunay, E. Khoury, A. Lazaridis, and S. Marc, "On the vulnerability of speaker verification to realistic voice spoofing," in *Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Virginia, USA, 2015, pp. 1-8.

Publications

International Journals

[1]. **Tanvina B. Patel** and Hemant A. Patil, "Significance of source-filter interaction for classification of natural *vs.* spoofed speech", accepted in *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification [Available Online: 15 March 2017].

[2]. **Tanvina B. Patel** and Hemant A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection", accepted in *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification [Available Online: 30 Dec. 2016].

International Conferences

[3]. Himanshu Bhavsar, **Tanvina B. Patel** and Hemant A. Patil, "Novel nonlinear prediction based features for spoofed speech detection", in *Proc. Annual Conf. of Int. Speech Comm. Assoc. (ISCA), INTERSPEECH*, San Francisco, USA, pp. 155-159, 8-12 Sept., 2016.

[4]. Meet Soni, **Tanvina B. Patel** and Hemant A. Patil, "Novel subband autoencoder features for detection of spoofed speech", in *Proc. Annual Conf. of Int. Speech Comm. Assoc. (ISCA), INTERSPEECH*, San Francisco, pp. 1820-1824, USA, 8-12 Sept., 2016.

[5]. Avni Rajpal, **Tanvina B. Patel**, Hardik B. Sailor, Maulik C. Madhavi, Hemant A. Patil and Hiroya Fujisaki, "Native language identification using spectral and source-based features", in *Proc. Annual Conf. of Int. Speech Comm. Assoc. (ISCA), INTERSPEECH*, San Francisco, USA, pp. 2383-2387, 8-12 Sept., 2016.

[6]. Deep Gandhi, **Tanvina B. Patel** and Hemant A. Patil, "A novel lowpass filtering-based approach for estimating strength of excitation from speech signal," in *Int. Conf. on Sig. Process. and Comm. (SPCOM)*, IISc, Bangalore, India, pp. 1-5, 12-15 June, 2016.

[7]. **Tanvina B. Patel** and Hemant A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (*SoE*) for spoofed speech detection" in *Proc. Int. Conf. Acoust., Speech and Signal Process., (ICASSP),* Shanghai, China, pp. 5105-5109, 20-25 March, 2016.

[8]. **Tanvina B. Patel** and Hemant A. Patil, "Analysis of natural and synthetic speech using Fujisaki model" in *Proc. Int. Conf. Acoust., Speech and Signal Process., (ICASSP)*, Shanghai, China, pp. 5250-5254, 20-25 March, 2016.

[9]. **Tanvina B. Patel** and Hemant A. Patil, "Combining evidences from Mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural *vs.* spoofed speech," in *Proc. Annual Conf. of Int. Speech Comm. Assoc. (ISCA), INTERSPEECH*, Dresden, Germany, pp. 2062-2066, 6-10 Sept., 2015.

[10]. Pramod Bachhav, Hemant A. Patil and **Tanvina B. Patel**, "A novel filtering based approach for epoch extraction," in *Proc. Int. Conf. Acoust., Speech and Signal Process., (ICASSP)*, Brisbane, Australia, pp. 4784-4788, 19-24 April, 2015.

[11]. Hemant A. Patil and **Tanvina B. Patel**, "Chaotic mixed excitation source for synthesis of speech signal," in *Proc. Annual Conf. of Int. Speech Comm. Assoc. (ISCA), INTERSPEECH*, Singapore, pp. 785-789, 14-18 Sept., 2014.

[12]. **Tanvina B. Patel** and Hemant A. Patil, "Novel approach for estimating length of the vocal folds using Fujisaki model," in *Proc. Int. Sym. on Chinese Spoken Lang. Process. (ISCSLP),* Singapore, pp. 308-312, 12-14 Sept., 2014.

[13]. Nirmesh Shah, Hemant Patil, Maulik Madhvi, Hardik Sailor and **Tanvina B. Patel**, "Deterministic annealing EM algorithm for developing Gujarati TTS system," in *Proc. Int. Sym. on Chinese Spoken Lang. Process. (ISCSLP),* Singapore, pp. 526-530, 12-14 Sept., 2014.

[14]. Hemant A. Patil and **Tanvina B. Patel**, "Nonlinear prediction of speech using Volterra-Wiener Series," in *Proc. Annual Conf. of Int. Speech Comm. Assoc. (ISCA), INTERSPEECH*, Lyon, France, pp. 1687-1691, 25-29 August, 2013.

[15]. Hemant A Patil, **Tanvina B. Patel**, Nirmesh J Shah, Hardik B Sailor, Raghava Krishnan, G. R. Kasthuri, T. Nagarajan, Lilly Christina, Naresh Kumar, Veera Raghavendra, S. P. Kishore, S R M Prasanna, Nagaraj Adiga, Sanasam Ranbir Singh, Konjengbam Anand, Pranaw Kumar, Bira Chandra Singh, S L Binil Kumar, T. G. Bhadran, T. Sajini, Arup Saha, Tulika Basu, K. Sreenivasa Rao, N. P. Narendra, Anil Kumar Sao, Rakesh Kumar, Pranhari Talukdar, Purnendu

Acharyaa, Somnath Chandra, Swaran Lata and Hema Murthy, "A syllable-based framework for unit-selection synthesis in *13* Indian languages," in *Proc. Oriental Int. Committee for the Co-Ordination and Standardization of Speech Databases and Assess. Techniques (COCOSDA) Conf*., Gurgaon, India, pp. 1-8, 25-27 Nov., 2013.

[16]. Hemant A. Patil, **Tanvina B. Patel**, Swati Talesara, Nirmesh Shah, Hardik Sailor, Bhavik Vachhani, Janaki Akhani, Bhargav Kankariya, Yashesh Gaur and Vibha Prajapati, "Algorithm for speech segmentation at syllable-level for text-to-speech synthesis system in Gujarati," in *Proc. Oriental Int. Committee for the Co-Ordination and Standardization of Speech Databases and Assess. Techniques (COCOSDA) Conf*, Gurgaon, India, pp. 1-7, 25-27 Nov., 2013.

[17]. Swati Talesara, Hemant A. Patil, **Tanvina B. Patel**, Hardik Sailor and Nirmesh Shah, "A novel Gaussian filter-based automatic labeling of speech data for TTS system in Gujarati language," in *Proc. Int. Conf. on Asian Lang. Process. (IALP)*, Urumqi, China, pp. 139-142, Aug., 2013.

Biography

Tanvina B. Patel received her Bachelor's B.E. degree in Electronics and Communication (E.C.) Engineering from Government Engineering College (GEC), Surat, in 2009 and M.E. degree in Communication Systems Engineering (C.S.E) from L. D. College of Engineering, Ahmedabad, Gujarat, in 2012. In July 2012, she joined the Doctoral program at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India.

From November 2016 to March 2017, she was a Research Assistant at Department of Electronics and Information Technology (DeitY) sponsored consortium project on "Speech Based Access of Agricultural Commodity Prices and Weather Information in 12 Indian Languages/Dialects (Automatic Speech Recognition (ASR) Consortium-Phase-II). From May 2012 to March 2016, at the same position she was associated with the consortium project entitled, "Development of Text-to-Speech (TTS) Synthesis Systems in Indian Languages-Phase-II" at DA-IICT, Gandhinagar.

Ms. Patel is a member of IEEE Signal Processing Society (SPS) and International Speech Communication Association (ISCA). She has received ISCA student grant to present research paper at INTERSPEECH 2013, Microsoft Research (MSR), India, travel grant to present research papers at INTERSPEECH 2014 and ISCSLP 2014 and IEEE SPS student travel grant to present papers at ICASSP 2016. Her research interest includes speech signal analysis, speech synthesis, and spoof speech detection for voice biometrics.