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].

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
Chapter 1: Introduction

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
Motivation for Spoof Detection Problem

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.
Chapter 1: *Introduction*

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.
Chapter 1: *Introduction*

**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.
Contributions from the Thesis

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 Liljencrantz-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 \( \hat{g}(t) \) 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
Organization of the Thesis

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 1: *Introduction*

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.