Doctoral theses at NTNU, 2021:394

Hareesh Mandalapu

Robust Algorithms for Audio-Visual Biometric Authentication

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of Science and Technology

Hareesh Mandalapu

Robust Algorithms for Audio-Visual Biometric Authentication

Thesis for the Degree of Philosophiae Doctor

Gjøvik, December 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

ISBN 978-82-326-6792-5 (printed ver.)
ISBN 978-82-326-5792-6 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2021:394

Printed by NTNU Grafisk senter

Declaration of Authorship

I, Hareesh Mandalapu, hereby declare that this thesis and the work presented in it are entirely my own. Where I have consulted the work of others, this is always clearly stated. Neither this nor a similar work has been presented to an examination committee elsewhere.

signature:

.....

(Hareesh Mandalapu)

Gjøvik, Date: 15/09/2021

Abstract

Biometric authentication has been a vital method of human identification across a wide range of applications. It is a quick and secure authorization process in applications such as border control and banking transactions. Recent advances in technology have triggered the use of embedded biometrics in smartphones and handheld devices. For reliable authentication, biometrics perform better when compared to traditional techniques such as passwords. Moreover, its advantages, such as permanence and uniqueness, have increased the growth of biometrics in everyday usage. However, there are certain limitations to biometric systems in providing optimal performance. These limitations play a crucial role in formulating artefacts to conceal or recreate the identity of individuals. Therefore, this raises concerns about the robustness of a biometric system and questions the accuracy of biometric recognition. In a real-world scenario, an end-user biometric system is protected from tampering by external sources. However, the only source of interaction is through the data capturing sensor. Thus, external artefacts, namely presentation attacks, have been a severe threat to biometric systems. On the other hand, the internal dependencies typically come from the limitations of the hardware and software deployed in the biometric verification workflow. Examples of internal dependencies vary from noise in biometric data to dependencies of biometric algorithms.

In this thesis, we have focused on improving the generalization of biometrics by working on some of the problems caused by presentation attacks and internal dependencies in biometrics. The key challenges in audio-visual biometrics were identified, and research objectives were designed for this thesis. The vulnerabilities in audio-visual biometrics are observed with the help of a thorough review of existing recognition and presentation attack detection methods. An exhaustive and comprehensive study along with a comparison and categorization of state-of-the-art methods have resulted in a novel dataset. The dataset includes different attributes, which provide the scope to perform extensive experiments to understand

dependencies and vulnerabilities. The thesis proposes a fusion of texture features based iris presentation attack detection algorithm, with results showing superior performance. Further, the cross-dataset experiments led to an empirical evaluation of vulnerabilities in iris biometrics due to presentation attacks.

In speaker recognition research, voice impersonation, language dependency and audio replay attacks cause high vulnerability. This thesis proposes a novel voice impersonation dataset with three different languages. The impact of voice impersonation as an attack is evaluated on the state-of-the-art speaker verification methods. Further, the speaker's language mismatch in the enrollment and testing steps of speaker verification is examined. The recent progress in smartphone usage is also reflected in high-quality speakers and microphones. A replay attack dataset is created with multiple smartphones as playback and recording devices, and vulnerability is examined.

The thesis examines the generalizability of biometric algorithms to improve the robustness of biometric recognition. The results from the proposed methods are evaluated with extensive experiments and detailed examinations of both publicly available databases and new datasets created in this work. In conclusion, the thesis proposes novel methods and approaches to examining the vulnerabilities in audio-visual biometrics, presentation attack detection (PAD) in iris and voice biometrics, the study of language dependency and audio replay attacks. The methods presented are valuable contributions to the research fields in developing robust smartphone biometric methods by addressing vulnerabilities from multiple sources.

Acknowledgement

I would like to thank my supervisors Professor Raghavendra Ramachandra, and Professor Christoph Busch for providing me with an incredible opportunity to pursue Ph.D. and constant support over the duration of this thesis. I would also like to thank all the institutes and research committee members from IARPA and SWAN research projects for funding my thesis. I thank the administration of department of information security and communication technology, NTNU, Gjøvik and university of applied sciences, Darmstadt for providing me with good research infrastructure. I want to express my gratitude to the co-authors of my publications Raghavendra Ramachandra, Christoph Busch, Aravind Reddy and others for their valuable collaboration. I would like to express my gratitude towards the members of dissertation committee for spending their time on reviewing my thesis and evaluation. I would like to thank my friends, colleagues and family members for their encouragement during tough times.

Contents

List of Tables

List of Figures

List of Abbreviations

APCER Attack Presentation Classification Error Rate

ASM Active Shape Model

AVSR Audio-Visual Speaker Recognition

BATL Biometrics Authentication with a Timeless Learner

BPCER Bona Fide Presentation Classification Error Rate

BSIF Binarized Statistical Image Features

CCA Canoncial Corelation Analysis

CHMM Coupled Hidden Markov Models

CNN Convolutional Neural Network

COTS Commercial off-the-shelf system

DET Detection Error Trade-off

DNN Deep Neural Network

EER Equal Error Rate

FMR False Match Rate

FNMR False Non-Match Rate

FOCS Face and Ocular Challenge Series

GMM Gaussian Mixture Models

IAPMR Impostor Attack Presentation Match Rate

IARPA Intelligence Advanced Research Projects Activity

IEC International Electrotechnical Commission

IKFD Incremental Kernel Fisherface discriminant

IMP IITD Multispectral Periocular

ISO International Organization for Standardization

LBP Local Binary Patterns

LSTM Long Short-Term Memory

M2VTS Multimodal Verification for Teleservices and Security Applications

NIR Near-infrared

NIST National Institute of Standards and Technology

PA Presentation Attack

PAD Presentation Attack Detection

PAI Presentation Attack Instrument

SOTA State-of-the-art

SVM Support Vector Machine

SWAN Secure Access Control over Wide Area Network

TIMID Texas Instruments and Massachusetts Institute of Technology

Part I

Overview

Chapter 1

Introduction

Automatic human identification has been a key process of authentication in modern day security systems. The use of biometrics is proved to be a quick and secure process of authorization in applications such as border control and banking transactions. The embedded biometrics have been used in smartphones and handheld devices due to the growth of technology [13]. Recognition algorithms utilize advanced sensors for biometric data capture used in human identification. For example, near-infrared cameras for capturing iris patterns and 3D dot projectors for optimal face recognition. Similarly, the smartphone has evolved with embedded biometric sensors to provide secure authentication in mobile applications. However, there are certain limitations to biometric systems in providing optimal performance. These limitations play a crucial role in formulating artefacts to conceal or recreate the identity of individuals. Therefore, this raises concerns about the robustness of a biometric system and questions the accuracy of biometric recognition.

This thesis focuses on developing novel approaches to improve the robustness of biometric systems in dealing with vulnerabilities and dependencies. The general dependencies of biometric systems come from sample quality, sensor specificity, behavior patterns. The use of a biometric system under different configurations introduces dependencies. For example, in smartphone biometrics, the data capturing process occurs under different lighting conditions, which adds unwanted noise. In this situation, a generalizable biometric system should take care of the problem of signal noise. A comprehensive survey of audio-visual biometrics is performed, and a multi-attribute smartphone biometric dataset is created to examine the problem of generalizability. The key aim of this dataset is to investigate the advantages of multimodal biometrics in dealing with the problem of robustness.

Further, from external sources, artefacts such as presentation attacks (PAs) attempt to hide or steal the identity of a target. A robust presentation attack detection (PAD) algorithm should prevent attacks from unknown or unseen artefacts. Therefore, a novel method for the detection of contact lens attacks in iris biometrics is proposed. However, database dependency is observed, and an empirical evaluation of texture-feature based iris PAD methods is carried out. In the same direction, the challenge of detecting voice impersonation in speaker recognition is studied by proposing a multilingual voice impersonation dataset. The language dependency experimented with cross-lingual speaker verification in four languages. Alongside this, smartphone audio replay attacks were created in ten different configurations, and vulnerability analysis performed. This thesis explores the different factors that challenge the robustness of audio-visual biometric authentication and proposes novel approaches/analyses to understand the problem of generalizability.

The thesis has been funded by the BATL and SWAN projects at Darmstadt University of Applied Sciences, Germany and the Norwegian University of Science and Technology (NTNU).

1.1 Overview of projects

1.1.1 BATL project

Biometric Authentication with a Timeless Learner (BATL) is part of Odin's Thor program funded by the Intelligence Advanced Research Projects Activity (IARPA) of the United States government to develop novel biometric technologies for presentation attack detection (PAD). The partners in this project are the Computer Science department of the University of Southern California (USC), the Idiap Research Institute (Switzerland), Darmstadt University of Applied Sciences (Germany), TREX Enterprises and Northrop Grumman. The goal of this project is to identify presentation attacks and ensure the subject is being correctly identified.

In cooperation with the Norwegian Biometrics Laboratory (NBL) at NTNU Gjøvik, the biometric research group at Hochschule Darmstadt works on presentation attack detection in iris biometrics. The team has developed a multimodal PAD at the USC Information Sciences Institute (USC ISI) to detect presentation attacks (PAs) in the face, iris and fingerprint modalities. The target of this PAD system is to perform a robust, accurate and timely detection of known and unknown (PAs). A set of novel sensors and machine learning techniques are employed to obtain PAD features and obtain the interoperability and generalizability of PAD algorithms. The PAD decisions from all three modules and unknown detectors are fused to accurately discriminate PAs, impostors, and identity concealers.

1.1.2 SWAN project

The Research Council of Norway funds the Secure Access Control over Wide Area Network (SWAN) project. The objective of the SWAN project is to promote research into and the development of a secure access control platform for mobile devices. The research methodology of the SWAN project is divided into four parts: Trustworthy biometrics, Privacy-preserving biometrics, Trustworthy transaction protocols and information fusion.

The enormous growth of smartphone technology has triggered a severe necessity for security protocols. Recent mobile devices come with expert hardware and adaptive software for many kinds of applications. Banking applications and identity verification systems have been using mobile devices given their high performance capabilities. In parallel, the threats to mobile devices have become apparent in the form of data hacks or illegal access. In these scenarios, biometrics can provide secure access to the devices with quick and easy usage. However, current mobile biometrics are prone to vulnerabilities such as presentation attacks. Therefore, trustworthy biometrics are a significant focus in the SWAN investigation of different sources of vulnerabilities. The biometric data contains sensitive information, and such data becoming available for misuse can lead to psychological and financial consequences. Novel privacy-preserving techniques are developed in the SWAN project using template protection methods. The biometric data collected is protected by following the privacy by design framework.

The trustworthy transaction protocols play a crucial role in financial transactions over communication channels. Web-based technologies are prone to malicious attacks originating from various devices connected to the network. In this regard, the SWAN project's advanced transaction protocols are developed to overcome the problems of harmful malware in online transactions. The biometric tem plate is stored on the client device to prevent data leakage thus dropping the disadvantages of central storage. Further, the SWAN project is developing a multimodal system employing more than one biometric characteristic for mobile banking applications. Depending on the cost of the transaction, a single biometric characteristic may not be enough in providing sufficient security. Therefore, a multimodal system is being developed with efficient biometric fusion at the feature, score, or decision level in the SWAN project.

1.2 Motivation and Problem Statement

The growth in technology and computational power has increased the amount of data processing to an enormous level. The digitally processed data contains multiple types of sensitive information, which is crucial to privacy issues. The protec-

tion of such data is a vital requirement in many applications. Therefore, a potential authorization process is employed in sensitive data processing applications. In recent decades, biometrics-based authentication has been an optimal way of person authentication. Biometric systems use unique, permanent and stable characteristics to authorize individuals to access sensitive information. The advantage of biometrics over traditional passwords or key cards is that biometrics is very quick and user friendly. However, the end-user biometric systems deployed come with multiple system dependencies and vulnerabilities that concern the robustness of human authentication. The system dependencies are formed because of the use of biometrics in a variety of environments. Different aspects such as capturing conditions, background noise, or human behavior may alter the performance of authorization. On the other hand, artefacts such as presentation attacks try to override the target human's identity with the help of manufactured biometric characteristics such as printed face, or recorded audio.

The primary motivation of this thesis is to address the challenges from the various factors that alter the consistent robustness of biometrics. The target of this work is to achieve the generalizability of biometrics under several real-world conditions. Generalizable biometrics should be able to display robustness under variable conditions. Therefore, employing such biometric systems in applications like mobile biometrics or smartphones would lead to trustworthy human authentication.

1.3 Research Objectives

The research objectives of this thesis are to study and propose the robustness of biometrics in the audio-visual domain in the scope of general and smartphone biometrics. The following research objectives are the target of this thesis.

1. To perform a comprehensive survey of audio-visual biometrics with an exhaustive study of all the aspects and support the development of an audio-visual dataset considering different attributes.
2. To develop novel presentation attack detection algorithms in iris biometrics using fusion of texture feature based information.
3. To study and experiment voice impersonation attack in speaker recognition with the help of a novel dataset.
4. To investigate the impact of change in behavioral patterns in automatic speaker verification.
5. To create a novel smartphone audio replay attacks dataset to examine different attack configurations.

1.4 Research Questions

The following research questions are framed upon the study of the background and identifying the problem statement.

1. Can the problem of vulnerability be reduced by using a multimodal recognition system? (Related chapters: 7, 6)

   Multimodal biometric systems are used to provide more accurate authentication than a single biometric cue [14]. Alongside this, research on anti-spoofing or presentation attack detection (PAD) methods focused on creating a special module in a biometric system [15]. Multimodal systems contain more than two biometric characteristics. Thus, it is useful to counteract spoofs by taking advantage of complementary information instead of creating an overhead through special modules. This could be achieved by studying the available audio-visual multimodal systems by detailed categorization and classification. This research question concerns anti-spoofing techniques using multimodal recognition systems and examines the existing recognition methods for multiple vulnerabilities.

2. Can PAD algorithms be generalized to unknown presentation attacks? (Related chapters: 8,9,10)

   One of the challenging problems in PAD is obtaining a generalization of the PAD algorithm. Given the dependencies of the developed PAD method on various attributes (e.g. sensor), unknown attacks cause a severe problem. This problem could be addressed by novel PAD methods that are robust against new kinds of presentation attacks. The proposed PAD methods should be tested for various real-world situations to examine their generalizability. The impact of presentation attacks and proposed PADs are evaluated over unknown scenarios and multiple dependencies. Further, unknown or least discussed presentation attacks are tested to impact biometric recognition and PAD performance.

3. Can the fusion of texture features be used in modelling generalizable PAD algorithms? (Related chapters: 8)

   The optimal fusion of different texture-based information obtained from the biometric data samples can be used for implementing PAD methods. Presentation attacks contain cues from the artefacts that are used to create them. This depends on the type of attack, the sensor used and the conditions of the data acquired. In the case of attacks where the artefact is highly similar to a bona fide, combination of texture features would provide better knowledge of attacks. This research question is about exploring quality

and texture-based features in severe presentation attacks to propose a novel PAD method. Contact lens attacks are prone to be challenging in detection. Therefore, image quality and a texture feature of periocular regions can be utilized here to identify the contact lens in an iris image. Further, the proposed method can be examined for generalizability.

4. Can using multimodal PAD algorithms be beneficial in overcoming the problems with vulnerabilities? (Related chapters: 7)

   A multimodal system comes with complementary biometric cues. The problem of presentation attacks can be addressed with the help of multimodal systems. For example, in an audio-visual system, an attack on the audio channel can be ignored with the help of a robust visual channel. Therefore, in detecting presentation attacks, multiple modalities would import additional characteristics for bona fide and artefact samples. This research question intends to examine the benefits of audio-visual systems by creating a set of novel presentation attacks. Multimodal presentation attacks such as synchronous replay attacks and synthesized attacks on individual cues are created in audio-visual domain. The complexities involved in implementing multimodal presentation attacks are investigated and, further, detecting attacks using the complementary data from bona fide samples.

1.5 Research Methodology

The research methodology of this thesis correlates with the research questions presented in the previous section. The following methodologies are planned to fill some gaps in this research domain by addressing the research questions. The key research topics of this thesis are presented in Figure 1.1.

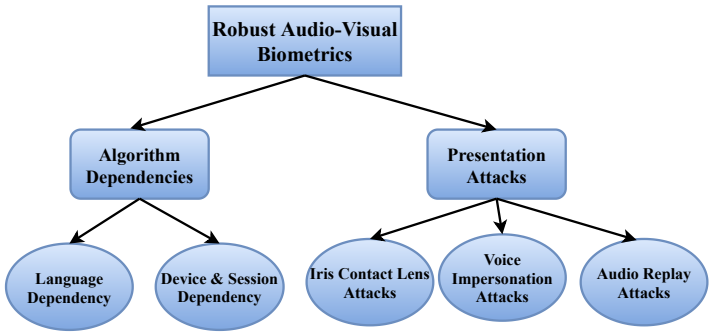


Figure 1.1: Research topics in this PhD program.

- **Comprehensive survey and novel audio-visual biometric dataset**

A comprehensive survey of audio-visual biometric recognition and presentation attack detection is performed. The key concepts of audio-visual biometrics, terminology and standards are explained. A detailed study of the datasets and benchmarking biometric algorithms is presented. The survey includes a classification and comparison of recognition and presentation attack detection (PAD) methods.

The drawbacks of the previous datasets in the audio-visual domain are identified. A novel dataset is created with 103 subjects in a smartphone environment, including multiple dimensions such as devices, sample noise and languages. The dataset is benchmarked with state-of-the-art biometric recognition algorithms. Two types of presentation attack are created in physical and logical access domains. The vulnerability of presentation attacks and the performance of baseline PAD methods are evaluated through extensive experiments.

- **Presentation attacks**

  Iris contact lens attack detection is a challenging problem in iris biometrics. A novel approach is proposed using an efficient fusion of image quality and texture features. The proposed method is tested on publicly available iris presentation attack databases. An empirical evaluation of existing state-of-the-art texture feature based iris PAD methods is performed, and the results are presented. The results show a consistent superiority compared to other texture feature methods, but dependency on the dataset is observed.

  Voice impersonation is the least discussed presentation attack in automatic speaker verification systems. A novel dataset of voice impersonation is created in three different languages using a crowd-sourcing approach. The impact of voice impersonation as a presentation attack is tested on state-of-the-art deep learning methods and baseline voice PAD methods. The dataset created is used to test voice impersonation attacks and the dependency of language.

- **Robustness Generalizability of biometric algorithms**

  Language dependency is examined using cross-lingual speaker verification where the language is different in training, enrollment and testing. Thorough experiments are performed on a publicly available smartphone dataset with four different cross-languages. The impact of language mismatch is observed under different classifiers using two speaker verification methods.

  A novel smartphone audio-visual dataset is examined for generalizability of biometric methods in three different scenarios: Inter-device, inter-session and inter-language. When the dependency arises, the biometric algorithms

display a drop in performance. Further, the problem of presentation attacks is also observed using two types of presentation attack.

A novel audio replay attack dataset across multiple smartphone configurations is developed. The record-playback configurations are carefully chosen to accommodate the impact of the bona fide data capture device. The attack data also contains multiple languages to observe the language mismatch problem in presentation attacks. The vulnerability of attacks is carried out by two methods: A state-of-the-art and a commercial-off-the-shelf method.

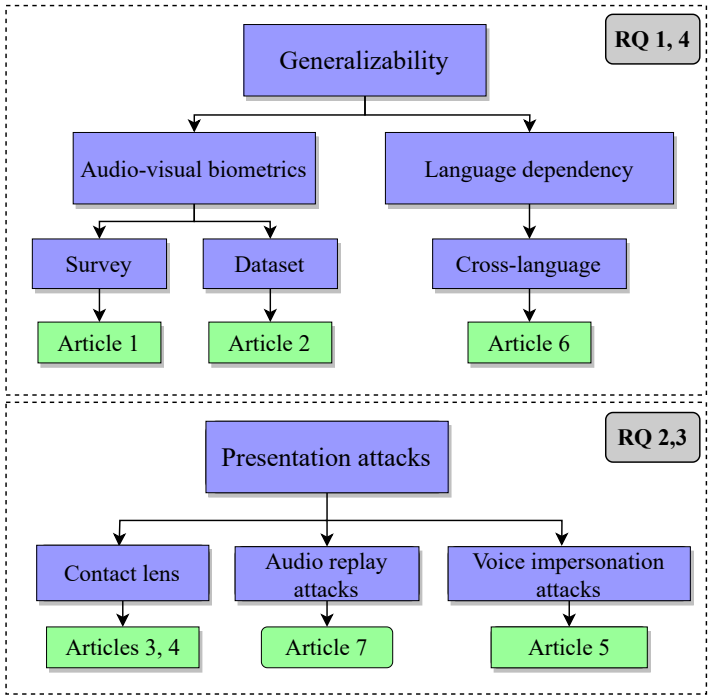


Figure 1.2: Research questions and corresponding published research articles.

1.5.1 Scope of the thesis

The scope of this thesis is to study and examine the dependencies from internal and external factors that challenge the robustness of audio-visual biometric systems. Further, with the help of the study, the focus is to develop novel techniques to address the problems caused by the dependencies, and presentation attacks in audio-visual biometrics. We have identified different scenarios where the performance of a biometric system can be affected. The vulnerability caused by these scenarios is examined with novel datasets over state-of-the-art biometric systems.

A comprehensive survey of audio-visual biometrics is carried out, and a multidimensional biometric dataset is created in a smartphone environment. The acquired dataset is benchmarked under different scenarios and novel presentation attacks. The next part of the thesis presents the challenging vulnerabilities caused by contact lenses in iris biometrics and voice impersonation in speaker recognition. A novel presentation attack detection method is proposed for contact lens detection, and a crowdsourcing based voice impersonation dataset is created to examine the problems of mimicry attacks. Further, the thesis presents the problem of language dependency through tested cross-lingual speaker recognition. The impact of audio replay attacks in smartphones under various configurations is observed in the last part of this thesis. The scope of the thesis is to provide an insight into the various challenges that impact the robustness of audio-visual biometrics. This thesis would attract the biometrics research community in modelling novel generalizable biometric systems.

1.6 List of Research Publications

1. Hareesh Mandalapu, Aravinda Reddy PN, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, SR Mahadeva Prasanna, and Christoph Busch. *"Audio-visual biometric recognition and presentation attack detection: A comprehensive survey."* IEEE Access, 9:37431–37455, 2021.

2. Hareesh Mandalapu, Aravinda Reddy PN, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, SR Mahadeva Prasanna, and Christoph Busch. *"Multilingual Audio-Visual Smartphone Dataset And Evaluation."* IEEE Access, doi: 10.1109/ACCESS.2021.3125485, 2021.

3. Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. *"Image quality and texture-based features for reliable textured contact lens detection."* In 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pages 587–594. IEEE, 2018.

4. Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. *"Empirical evaluation of texture-based print and contact lens iris presentation attack detection methods."* In Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications, pages 7–14, 2019.

5. Hareesh Mandalapu, Raghavendra Ramachandra and Christoph Busch. *"Multilingual voice impersonation dataset and evaluation."* In Proceeding of the 3rd International Conference on Intelligent Technologies and Applications (INTAP). Springer, 2020.

6. Hareesh Mandalapu, Thomas Møller Elbo, Raghavendra Ramachandra, and

Christoph Busch. *"Cross-lingual speaker verification: Evaluation on x-vector method."* In Proceeding of the 3rd International Conference on Intelligent Technologies and Applications (INTAP). Springer, 2020.

7. Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. *"Smartphone audio replay attacks dataset."* In 2021 9th IEEE International Workshop on Biometrics and Forensics (IWBF). IEEE, 2021.

1.7 Thesis Outline

This thesis is divided into three parts. Part I contains an overview of the thesis, which presents the introduction, background, a summary of published articles, conclusion and future work. Chapter 1 describes the BATL and SWAN projects, the motivation for this thesis, problem statement, research methodology and list of research articles. Chapter 2 discusses the background and related work on the key topics and terminology that is used in the other parts of this thesis. Chapter 3 summarizes the research articles which are part of this thesis. The conclusion and future work are discussed in Chapter 4 and 5 respectively.

Part II presents the research articles that are related to the research methodology of the thesis. Chapter 6 presents the paper on the comprehensive survey of audio-visual biometric recognition and presentation attack detection. In Chapter 7, a novel audio-visual biometric dataset created in a smartphone environment is described along with benchmarking experiments and results. Chapter 8 presents the paper on the novel approach to contact lens detection in iris biometrics with extensive experiments. In the this direction, an empirical evaluation of texture-based iris presentation attack detection methods is presented in Chapter 9. Moving further in the voice biometrics direction, Chapter 10 presents the voice impersonation dataset collected to evaluate the impact of mimicry attacks on state-of-the-art speaker recognition methods. Chapter 11 describes the cross-lingual speaker verification experiments on the X-vector method to observe the impact of language dependency. Finally, the impact of audio replay attacks in smartphone biometrics is examined, and the results are presented in Chapter 12.

Part III presents the appendix of the thesis with a discussion of mobile applications in capturing smartphone biometric data in 7.

Chapter 2

Background and Related Work

In this chapter, we present the fundamental concepts of biometrics, including physiological and behavioral biometrics in Section 2.1 with a discussion of the modalities used in this thesis. Then, we explain the topic of multimodal biometrics with emphasis on audio-visual biometrics in Section 2.2. The problem of generalizable biometrics with the topics of presentation attacks and algorithm dependencies is explained in Section 2.3.

2.1 Biometrics

Human identification has become a key feature of modern-day authorization. A human can be identified in three different ways: *what you have*, *what you know* and *what you are*. Key cards or identity cards come into the category of *what you have* whereas passwords or PIN codes fall under *what you know*. The third category *what you are* is the unique characteristics of which an individual consists. This type of authentication does not require carrying any extra item or remembering patterns. The unique characteristics every human possesses are biometrics. Biometrics is defined as the "*automatic recognition of individuals based on their behavioral or biological characteristics*" [16]. Biometric identification is more beneficial than the other two ways given properties such as uniqueness, permanence and user-friendliness.

A biometric recognition system captures the unique characteristics of an individual and performs signal processing to compare them with the registered signal in the database [13]. Depending on the type of characteristics, a biometric system uses different sensors in capturing. Similarly, various signal processing steps and comparison methodologies are employed in the biometric recognition process. Biometric cues are divided into two types: physiological and behavioral. The biometric

cues used in the recognition methods can be more than one modality: Unimodal and multimodal biometrics. More about the types of biometric characteristics and systems is explained in the following sections.

2.1.1 Physiological Biometrics

Physiological biometrics are biometric characteristics, including biological or physiological features of the human. Typical physiological biometrics are fingerprints, the face and iris biometrics. In general, physiological biometrics are captured in a single shot scenario. Therefore, physiological biometrics have become easy to use and are deployed in a wide range of applications.

Face Recognition

Face recognition is the process of identifying a person with the unique properties of a person's face. Face recognition has evolved into an active biometrics research domain given its advances in capturing devices. Over the years, advanced face biometric methods have been proposed that display near-zero inaccuracy. Face image representations use texture-based features, animation-based features and, recently, deep learning features. Texture-based features are computed using filters and used for comparing facial images. Local Binary Patterns (LBP) is one of the popular texture features that display consistent performance in face recognition [17]. Other texture features include the Histogram of Gradients (HOG) [18], Gabor filters [19] and Haar filters [20]. Animation-based face features use the shape and appearance of a facial region in an image with the help of active shape models [21]. The animation features utilize high-level features such as the lip-contour region and have the advantages of light sensitivity and rotation [1]. Typical face animation features use a series of frames in a video to obtain optical flow, and motion blur [22] [23]. Deep learning methods take over face recognition research by allowing generalizable and robust biometric algorithms. FaceNet is a CNN that outputs face embeddings using triplet loss and is efficient in identifying similar faces [24]. ArcFace features are proposed for face biometrics with a higher level of discrimination by emphasizing the loss function [25]. FaceNet and ArcFace methods display consistent performance over different face biometrics databases.

Iris Recognition

The human iris has unique characteristics which are widely used for biometric recognition. The benefits of the iris over the face fingerprints are that it is a more protected organ and unique even in monozygotic twins [26]. Although state-of-the-art iris recognition systems require special sensors such as near-infrared cameras, it is submitted that iris biometrics is more robust and accurate than face- or fingerprint-based biometrics [27] [28].

However, iris recognition is also prone to vulnerabilities such presentation attacks [3].

2.1.2 Behavioural Biometrics

Behavioral biometrics are characteristics of human behavior over a short period. Popular behavioral biometrics are the voice, gait and keystroke dynamics. The capturing process for behavioral biometrics happens over a short time, ranging from a few seconds to minutes. Behavioral biometric recognition analyzes the patterns in a person's behavior and identifies the unique properties for recognizing a person. Time-series data in human behavior also recognizes other properties such as age, gender, and emotion.

2.1.3 Speaker Recognition

Automatic speaker verification (ASV) is a process of identifying a person based on speech patterns. Speaker recognition has attracted attention given its ability to authenticate remotely via telecommunication. The acoustic feature extraction from speech reflects the uniqueness of a speaker and also contains behavioral patterns. The cepstral features of an audio signal are widely used in speaker verification. Mel-frequency cepstral coefficients (MFCC) are popular acoustic features that are based on auditory perceptions [29]. It has been suggested that MFCCs represent the human perception of voice more accurately by suppressing minor variations in higher frequency bands. In ASV methods, MFCCs that are used along with Gaussian mixture models (GMMs) show superior results across different applications. I-vectors are low dimensional representations of a speech sample computed using MFCCs using Joint factor analysis (JFA) [30]. I-vectors model channel effects and information about the speakers in a vector representation. Probabilistic linear discriminant analysis (PLDA) [31] displayed better performance in training speaker models using i-vectors. The deep learning methods have evolved in representing a voice sample using acoustic features or raw audio. X-vectors are fixed dimensional deep neural network audio features used to differentiate speakers [32]. The deep learning model used in this approach is a feed-forward neural network on cepstral features [33].

2.2 Multimodal Biometrics

Biometric systems using only one biometric cue (unimodal system) have several problems given their limited data. Although a unimodal system can use multiple classifiers to perform biometric verification, the data obtained from the sensor can be problematic. Therefore, multimodal biometrics has become a popular research direction [14]. In general, multimodal biometric systems use more than one type of biometrics and employ a fusion approach to make a decision on recognition.

Multimodal systems have advantages such as additional information to overcome the problem of noisy data in one biometrics. Biometric identification systems with multiple cues are employed in many applications [34]. Many biometric databases provide multimodal data to encourage research in biometric fusion approaches [10], [8], [12].

2.2.1 Audio-Visual Biometrics

Audio-visual biometrics have attracted interest given their unique properties and advantages over multimodal biometrics. Multimodal biometrics may contain different biometric data and use different classification approaches to identify a person. Therefore, using multimodal systems introduces new problems such as capturing time, processing overhead and design difficulties. Audio-visual biometrics use a single capture system of a talking face and combine complementary correlated information, unlike multimodal systems. Alongside unimodal biometric challenges, audio-visual speaker recognition challenges are also attracting research [35]. In smartphones, audio-visual biometrics can be deployed as modern mobile devices contain video cameras and microphones. Mobile biometric applications like e-commerce and mobile payments can take advantage of audio-visual biometrics to provide high-level authentication at less cost [36].

2.3 Generalization problem

The problem of generalization is the problem of the inconsistent performance of a biometric system across different setups. Biometric systems are impacted by two different types of factor, namely internal dependencies and external artefacts. The internal dependencies are the factors included while developing a biometric system and limit the performance in other scenarios, e.g. data noise, behavioral patterns, capturing devices etc. On the other hand, external artefacts are attacks on deployed biometric systems in order the alter their performance, e.g. presentation attacks. The generalization problem challenges the robustness of biometric algorithms.

2.3.1 Algorithm Dependencies

The internal dependencies of a biometric system are accumulated from several sources. Popular dependencies are the type of biometric data used in development, the variance of capturing devices, and changes in the behavioral patterns of subjects [21]. In smartphone biometrics, biometric data varies because the data capture is not under controlled situations. Also, the developed algorithm is deployed in different devices, and the biometric sensor (camera) introduces new properties to the data. Behavioral patterns like changes in language or text impact the performance of voice-based biometrics.

2.3.2 Presentation Attacks

There are several vulnerable points in a biometric system caused by external artefacts, as shown in Figure 2.1. According to ISO/IEC standards [5], *presentation attacks* are defined as the presentation of a biometric capture subsystem to interfere with the operation of the biometric system. The artefact used in this process is called a Presentation Attack Instrument (PAI). There are two types of presentation attack: An active impostor presentation attack where an attacker tries to be recognized as a different subject, and a concealer presentation attack where the attacker avoids being recognized as a subject in the system. The increase in vulnerability caused by presentation attacks has given rise to a new module of presentation attack detection (PAD) in biometric systems. Presentation Attack Detection (PAD) is the identification of presentation attacks to be classified, particularized, and communicated for decision-making and performance analysis [37]. PAD is also termed anti-spoofing, or liveness detection in the literature [15].

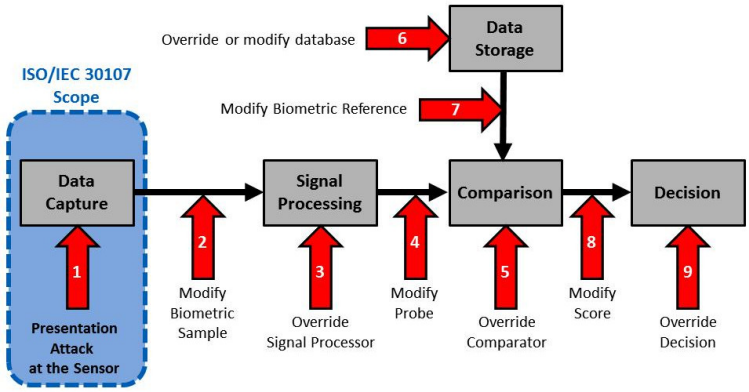


Figure 2.1: Presentaion Attacks on a biometric system [5].

In audio-visual biometrics, presentation attacks are generally performed on audio or video capturing sensors or both. In audio channels, presentation attacks are voice impersonation, audio replay, voice conversion, and speech synthesis [21]. In video channels, printed images, display presentations, synthesized signals (DeepFake or artificial face) and 3D masks [38], [28]. Alongside this, audio-visual replay attacks and digital audio-visual attacks are created using display-speaker setup and face-speech synthesis, respectively [39].

Chapter 3

Summary of Published Articles

In this chapter, we summarize the research articles published throughout this PhD program. The following sections present a brief overview of each article with an introduction, motivation, and research findings. The topics shown in Figure 1.1 and research questions in Section 1.4 correspond to the papers discussed in this chapter.

3.1 Article 1: Audio-visual biometric recognition and presentation attack detection: A comprehensive survey

Multimodal biometrics have attracted research attention given the scope for utilizing multiple sensors for biometric recognition. Among these, audio-visual biometrics are discussed in many works for their advantages, such as complementary and correlated biometric cues. The advantage of audio-visual biometrics over other multimodal biometrics is that audio-visual biometrics can be acquired in a single capture and contain additional correlated information. Therefore, the growth of audio-visual biometrics has seen a constant growth in research.

A detailed survey on audio-visual (AV) biometrics and presentation attack detec tion methods is carried out in this paper. The paper introduces the topic of multimodal biometrics and the category of audio-visual biometrics with general concepts and related work in the literature. The terminology is provided according to the ISO/IEC standards [16]. The feature extraction methods in AV biometrics' audio and visual domains are explained in detail with classifications. The next section describes the fusion approaches used to combine efficiently the audio and video domains for biometric recognition. AV biometric databases are created in different domains such as smartphones, handheld devices and high-tech sensors. A thorough study of the AV databases is performed in this paper with a description

of databases and the best performing biometric method. Sample biometric images are also presented to provide an insight into the databases.

One of the advantages of AV biometrics is the complementary information present in the sample. This information is often utilized in identifying presentation attacks or forgery attacks. In this study, we have presented the features used in attack detection in AV biometrics with categorizations. The summary of the different features, databases and performance of attack detection methods is presented in a table. Further, the challenges and open questions in the field of AV biometrics are discussed. The main challenges include databases, AV biometrics in smart devices and performance evaluation protocols. The identified challenges would provide scope for valuable research on AV biometrics in the future.

3.2 Article 2: Multilingual Audio-Visual Smartphone Dataset And Evaluation

Smartphone biometrics has evolved into critical privacy and data security applications in daily life such as mobile banking and digital identity. Smartphone manufacturers embed additional sensors into devices to provide accurate authentication. However, the biometric system dependencies and external vulnerabilities restrict the robustness of biometric recognition. The well-known dependencies are signal noise, changes in behavior and channel variability. The external vulnerabilities include presentation artefacts that can be used to attack or conceal the identity of the target person. The wide range of devices, capturing conditions and growing artefacts impact the generalizable properties of biometric algorithms. In this regard, multimodal biometrics have come into play to include complementary information on different biometric cues. More importantly, audio-visual biometrics come with correlated biometric information to deal with dependencies and attacks.

In this paper, we have created a novel multilingual audio-visual smartphone (MAVS) dataset that accommodates the research scope for examining the generalizable properties of biometrics in smartphones. The main focus of this paper is to evaluate the impact of dependencies and attacks on the start-of-the-art algorithms on smartphone biometric data. Therefore, we created the dataset, which includes multiple sessions with variable lighting and noise, multiple smartphone devices and multiple languages. Further, we have created presentation attacks in two directions, namely physical and logical access.

Extensive experiments were performed in different scenarios observing the problems of biometric algorithms, and the results are presented in detail. The two different types of experiment are designed to examine the robustness of biometric algorithms. The first type of experiment includes internal dependencies such as

signal noise, capture device and audio language. The second experiments verify the impact of presentation attacks by checking vulnerability and attack detection methods. The results are presented in ISO/IEC standards, and comparisons are made to check the inter-device, inter-session and inter-language situations. The presentation attack detection methods are taken from the baseline attack detection challenges [40], [12].

The novel dataset proposed in this work can be utilized to address several challenges in audio-visual biometric research. Developing generalizable biometric algorithms across a wide variety of smartphones requires a dataset with several attributes discussed in this work. The key attributes in this dataset include the biometric data with multiple languages, devices and sessions. The multidimensional dataset has the capacity to research and propose robust biometric algorithms in a smartphone environment. The drawbacks in current recognition systems can be the subject of an experiment, and novel algorithms can be implemented with the help of observations. Further, the protocols used in creating the dataset and making presentation attacks can be used as benchmarks for creating a newly updated dataset.

3.3 Article 3: Image Quality and Texture-Based Features for Reliable Textured Contact Lens Detection.

Presentation attacks in iris biometrics cause a serious vulnerability despite being a unique and stable biometric modality [26]. The vulnerability can be in the form of concealing iris identity or an attack on an already enrolled iris. Artefacts such as printed iris images, or electronic display attacks can be detected with advanced censors such as near-infrared cameras. However, textured/patterned contact lens attacks show a significant problem for iris-based recognition [3]. The difficulty in manually detecting a contact lens is that it covers the iris region and moves along with the eye movements. The existing methods in this direction dealt with contact lens attacks passively, not taking into account different lens species and capturing devices. Therefore, we have proposed in this paper a novel contact lens detection method using a weighted fusion approach of image quality and a texture feature. The proposed method is tested for generalizability, and the results are compared with state-of-the-art algorithms.

The proposed method uses two features of a periocular image, namely BRISQUE and BSIF. Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) features are a statistic-based distortion-generic image quality assessment (IQA) model that provides a measure of image quality [41]. This feature includes the quality of the bona fide iris and distinguishes it from contact lens attacks. Binarized Statistical Image Features (BSIF) are texture-based features that are known to be

more suitable for detecting patterned/textured contact lens [42]. The two features are used to train two Spectral Regression Kernel Discriminant Analysis (SRKDA) classifiers independently [43]. The obtained test scores from two classifiers are combined using a weighted score-level fusion. The Fisher Discriminant Ratio (FDR) [44] is employed to obtain the weights on two different features.

Experiments were performed on multiple datasets publicly available from the LivDetIris 2017 challenge [45]. Results indicate superior performance over other texture feature-based attack detection methods. The results are consistent across different types of contact lens used in different datasets. However, testing for the generalizability of cross-sensor experiments, the proposed method failed to display similar performance across all datasets. The cross-sensor experiments were carried out by limiting the types of attack. The training data is taken from one of the datasets and tested on all other datasets. This inconsistency in performance leads to questioning the robustness of the presentation attack detection algorithms.

3.4 Article 4: Empirical Evaluation of Texture-Based Print and Contact Lens Iris Presentation Attack Detection Methods.

The robustness of iris presentation attack detection algorithms is heavily affected given the lack of generalizability. The state-of-the-art methods display optimal performances in limited conditions. When tested for different varieties of real-world scenarios, they often failed to perform optimally. Studies were performed on the assessment of iris presentation attack detection methods with categorization and taxonomy [46], [28]. In this paper, we empirically evaluated the well-known texture feature based on iris PAD methods in multiple scenarios. Texture- based features are one of the most popular categories of features used in iris PAD. The motivation behind this paper is to understand the behavior of PAD methods towards generalizability.

The experiments are designed to include different types of dependencies like unknown attack, unknown device and all combined. We have chosen four different types of dataset with two different types of attack. Five different texture feature-based methods were carefully chosen for the evaluation. The two types of attack used in this paper are two different presentation attack species: Print attacks and contact lens attacks. The five PAD methods utilize features such as LBP, BSIF, CAQP, BRISQUE and PHOG with the classifiers SVM and SRKDA.

The key observations from this paper are that no single method performs better in all the different scenarios. However, it is important to note that the efficient fusion of high performing texture features can deliver a better performing attack detection method. The performed evaluations provide an insight into the impact

of different scenarios that presentation attacks can be carried out in and how PAD methods perform in such situations. Future work in this direction suggests that robust biometric algorithms are required to consider various forms of vulnerability. Thus, the consistent performance of biometric recognition can be assured.

3.5 Article 5: Multilingual voice impersonation dataset and evaluation.

Behavioral biometrics are unique traits of characteristics that are used to identify humans. Unlike physiological biometrics, behavioral biometrics are an information series over a period of time. The behavior of humans such as speaking, walking are used as biometrics. Among these, speaker recognition has been prominently used for biometrics. Along with the progress of behavioral biometrics, vulnerabilities due to presentation attacks have also evolved. In speaker recognition, attacks on biometrics include voice impersonation, audio replay, voice conversion and speech synthesis. Voice impersonation is a physical access attack that does not include any equipment. Although voice impersonation is an obvious method of presentation attack, it is the least discussed in the research domain.

We have identified the problems existing in examining voice impersonation as a presentation attack. This paper created a voice impersonation dataset using a publicly available source, namely YouTube. We chose three different languages and accumulated popular impersonators through manual inspection. The speech samples from YouTube videos were created in a way similar to VoxCeleb datasets [47]. The audio samples with overlapped voices, background noise and dominating music were omitted. For each language, samples from fifteen different speakers were acquired with both bona fide and impersonation data. The protocols for creating this dataset are publicly available for research purposes.

We have examined the state-of-the-art biometric methods for the vulnerability of voice impersonation. With the help of pre-trained models from Kaldi, we evaluated I-Vector and X-vector based automatic speaker verification methods. The results show the considerable impact of impersonation attacks in all three languages. Although the vulnerability varies from language to language, the voice impersonation attack samples can match with the bona fide samples. The observations show the problem of impersonation attacks and trigger a need for subject-specific speaker verification systems independent of languages.

3.6 Article 6: Cross-lingual speaker verification: Evaluation on x-vector method.

Biometric algorithms are used across the world in a wide variety of applications. An end-user behavioral biometric system can be affected by a change in behavior or the subject. In speaker recognition, the difference in the language of the audio samples in enrollment and testing can alter the performance. In text-independent speaker recognition, language dependency has emerged as one of the key problems. This dependency on language reduces the robustness of speaker recognition systems [48]. Although many language-independent methods were proposed, the dependency of language is not adequately studied [49].

In this paper, we performed experiments on cross-lingual speaker verification using a smartphone dataset. The state-of-the-art deep neural networks approach called the x-vector method is tested for language dependency. Multiple language audio data from different sessions are examined in the experiments. The mismatch of languages during training, enrollment and testing were observed in different experiments. We chose two pre-trained models trained on NIST-SREI6, and VoxCeleb [50] datasets. The NIST-SREI6 training model contains two types of classifier 1. OOD PLDA is an out-of-domain model trained on data other than SRE16. 2. ADT PLDA is an in-domain PLDA that is adapted to the major partition of SRE16. We used the SWAN dataset that contains four different languages and five different sessions for enrollment and testing.

We carried out two experiments dependent on the training datasets. The first experiment used two types of PLDAs in NIST-SREl6 models to check the cross-lingual speaker verification of the X-Vector method. It was clearly observed that the language mismatch impacted the performance indicated by the drop in EER. The second experiment used the VoxCeleb model with only one PLDA. The decrease in performance due to language mismatch is high in this experiment. Also, the speaker recognition accuracy is low compared to the SRE16 model. The main reason for this behavior is the huge variance in data from the VoxCeleb datasets. Although both the datasets contain multiple languages and the X-vector method uses deep learning for training, the cross-lingual speaker recognition is limited.

3.7 Article 7: Smartphone audio replay attacks dataset.

Audio replay attacks are well-known presentation attacks on automatic speaker verification systems. The advances in high-quality speakers enabled accuracy in the replay of the voice in an audio sample. The replay attacks can be performed with the help of a digital copy of the target speaker audio. The growth in smartphone usage has increased the vulnerability of biometric systems embedded in

smartphones towards replay presentation attacks. Alongside smartphones, speakers have been growing to playback more frequencies present in human speech. The dependencies introduced by the devices and languages are other key factors that can add to replay attacks. From the discussion above, it is necessary to observe the impact of replay attacks and the perspective of device and language dependencies.

In this paper, we examined multiple configurations of smartphones in creating record-playback combinations and verified the vulnerability of two speaker verification systems: State-of-the-art and commercial-off-the-shelf. Further, we also performed experiments on baseline attack detection methods and presented the results. The bona fide audio data is extracted from session 1 audio-visual biometric data from the SWAN dataset [12]. The bona fide data was recorded using iPhone 6s in four different languages from 50 subjects, each speaking four sentences. We opted for five other smartphones to create ten different replay attack configurations. We created protocols for evaluating smartphone replay attacks in four languages and attack recording devices being the same or different from bona fide devices. The SWAN dataset also provides replay attack data which is used as training data in this paper.

The replay attack dataset created in this paper was tested over the X-vector-based state-of-the-art and VeriSpeak commercial methods. All the attack configurations display high vulnerability where most of the attacks are matched to the bona fide samples. Although the commercial method displays slightly less vulnerability than the state-of-the-art method, the high match rates are consistent. Similarly, when the bona fide recording device and attack recording device are the same, the vulnerability caused by the attacks is slightly decreased. When it comes to dependencies by languages and devices, there is no considerable difference in attack performance. This shows the problem of replay attacks independent of languages and devices among different record playback settings.

Further, the baseline attack detection methods of ASVSpoof 2019 [40] were examined for attack detection. It was observed that when the bona fide and attack recording device is the same, the PAD performed better. Also, there is a correlation between the device manufacturers in the PAD performance. When the devices in playback and recording are from the same manufacturers, the PAD displays better performance. This paper provided an insight into the audio replay attacks created using smartphones and their impact on smartphone biometrics.

Chapter 4

Conclusions

The main aim of this thesis is to develop state-of-the-art approaches to meet the research objectives and answer the research questions mentioned in Sections 1.3, 1.4, respectively. The research topics investigated in this thesis are divided into two parts, i.e. algorithm dependencies that affect generalizability and presentatin attacks (see Figure 1.1). Various research problems are identified in these fields and are used to frame four research questions. Seven research articles are included in this thesis that attempt to address the research questions. The thesis starts with a comprehensive review of audio-visual biometrics, including databases and feature extraction methods in recognition and presentation attack detection. The key output of this paper is the challenges and open questions identified in the current research.

Further, a novel audio-visual dataset was created in a smartphone environment that accommodates multiple attributes such as three languages, five devices and three ses sions with variable background noise and lighting. The dataset also created two types of novel presentation attack in physical access and logical access domains. We have evaluated the dependency of language in a cross-language speaker recognition scenario on state-of-the-art methods. Thus, the thesis provides valuable research on biometric algorithm dependencies and presentation attacks in audio-visual biometrics.

The contact lens attacks in iris biometrics pose a challenging problem in the robustness of iris based human recognition. In this direction, a novel presentation attack detection (PAD) algorithm is proposed to identify texture/patterned contact lenses. However, the generalizability of the proposed methods is not achieved in the case of cross-dataset evaluation. Therefore, an empirical evaluation is performed to observe PAD performance in different scenarios. The results provided

an insight into the behavior of PAD methods in real-world situations. Another least discussed presentation attack is voice impersonation in speaker recognition. Thus, this thesis chose the voice impersonation topic to answer our research question about unknown/future presentation attacks. A novel voice impersonation dataset was created in three languages and evaluated for vulnerability in automatic speaker verification methods. In the direction of presentation attacks, the smartphone audio replay attack dataset was created to observe the generalizability of attacks over different playback-record configurations and languages. Extensive vulnerability analysis and attack detection experiments demonstrated a replay attack's impact on speaker recognition performance.

The thesis achieves the research objectives through the collection of research articles. The research questions are answered with novel approaches, datasets, and experiments carried out throughout this thesis. The research methodologies, datasets and proposed approaches contribute to Robust Algorithms for Audio-Visual Biometric Authentication.

Chapter 5

Future Work

The robustness of audio-visual biometrics has been examined, and novel methods proposed in this thesis. The PAD method for iris contact lens attacks, the impact of voice impersonation and audio replay attacks in smartphones have been evaluated using novel datasets and detailed experiments. The following sections present the scope for future work based on the research work for this thesis.

5.1 Generalizability of Biometrics

Generalizable biometrics is a major requirement given the growth of embedded biometrics in many fields. In this thesis, we have examined the generalizability of audio-visual biometrics under multiple dependencies and algorithms. The major dependencies examined in the thesis are the type of dataset, capturing device, signal noise and language. Future research should identify other key problems and propose protocols to evaluate them for a robust biometrics. In this direction, advanced methods can be proposed to overcome the problems caused by the dependencies. The embedded biometrics are tested in multiple scenarios for the problem of generalizability. Alongside this, the presentation attacks have been evolving to conceal or steal the identity of the target subjects. Robust PAD algorithms for unknown and future presentation attacks are implemented in the future scope of this thesis.

5.2 Audio-Visual Biometrics

The advantages of audio-visual biometrics have been discussed exclusively in this thesis. The benchmarking experiments on the novel smartphone dataset will assist in proposing state-of-the-art multimodal biometric systems. Therefore, in the future scope of this thesis, complementary biometric cues in audio-visual biometrics

could be utilized to propose robust biometric algorithms. PAD methods can make use of audio-visual biometric synchrony and detect presentation attacks without the need for a dedicated PAD module for each cue. The visual speech or talking biometric face characterization would improve recognition robustness and prevent presentation attacks.

5.3 Presentation Attack Detection

This thesis used fusion of texture features obtained from the periocular image to propose an iris contact lens PAD method. The impact of voice impersonation and smartphone audio replay attacks on automatic speaker verification were examined through extensive experiments. Presentation attacks in unconstrained situations and most vulnerable scenarios are the key attributes affecting biometric systems. Deep learning approaches could model the liveness of a biometric sample and opt-out artefacts. Therefore, by precise modelling of bona fide samples, a presentation attack can be ruled out. In this direction, future work could investigate new approaches to create presentation attacks and propose PAD schemes to detect them.

Part II

Published Articles

Chapter 6

Article 1: Audio-visual biometric recognition and presentation attack detection: A comprehensive survey

Hareesh Mandalapu, Aravinda Reddy PN, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, SR Mahadeva Prasanna, and Christoph Busch. Audio-visual biometric recognition and presentation attack detection: A comprehensive survey. *IEEE Access*, 9:37431–37455, 2021.

6.1 Abstract

Biometric recognition is a trending technology that uses unique characteristics data to identify or verify/authenticate security applications. Amidst the classically used biometrics, voice and face attributes are the most propitious for prevalent applications in day-to-day life because they are easy to obtain through restrained and user-friendly procedures. The pervasiveness of low-cost audio and face capture sensors in smartphones, laptops, and tablets has made the advantage of voice and face biometrics more exceptional when compared to other biometrics. For many years, acoustic information alone has been a great success in automatic speaker verification applications. Meantime, the last decade or two has also witnessed a remarkable ascent in face recognition technologies. Nonetheless, in adverse unconstrained environments, neither of these techniques achieves optimal performance. Since audio-visual information carries correlated and complementary information, integrating them into one recognition system can increase the system's

performance. The vulnerability of biometrics towards presentation attacks and audio-visual data usage for the detection of such attacks is also a hot topic of research. This paper made a comprehensive survey on existing state-of-the-art audio-visual recognition techniques, publicly available databases for benchmarking, and Presentation Attack Detection (PAD) algorithms. Further, a detailed discussion on challenges and open problems is presented in this field of biometrics.

6.2 Introduction

Biometric technology is swiftly gaining popularity and has become a crucial part of day-to-day life. A biometric system aims to recognize a data subject based on their physiological or behavioral characteristics [51]. Recognition systems are based on biometric characteristics, such as DNA, face, iris, finger vein, fingerprint, keystroke, voice, and gait. Several factors are considered while designing and applying biometrics: accuracy to authentication, robustness to spoof or impostor attacks, user acceptance, and cost of capture sensors. Amidst these factors, user acceptance and sensor cost are the primary hindrances that thwart highly accurate and robust biometrics.

The authentication system that uses a single biometric cue such as speech or face is called a unimodal system. The biometric cue can use more than one classifier and employ a fusion approach to perform recognition. Nonetheless, the captured biometric cue may be of low quality due to variations in pose, illuminations, background noise, and low spatial and temporal resolution of the video. This problem is addressed by using multiple biometric modalities for authentication [14]. Deploying multimodal data introduces other problems like multiple captures, processing time, and design overhead. The vulnerabilities present by unimodal biometrics may also exist in a multimodal system. The audio-visual biometrics took multimodal biometrics to another better level by taking advantage of complimentary biometric information present between voice and face cues. In analogy, voice and face biometrics are most user-friendly and cost-effective as they allow capturing the multi-biometrics in a single capture using low-cost sensors (e. g., smartphone camera). These points made audio-visual biometrics an exciting topic of research in field of multimodal biometrics.

Audio-visual biometrics has gained interest among biometric researchers both in academics and in industry. As a result there are an ample amount of literature available [51, 52, 53, 54], publicly available databases [55, 8, 56, 57, 58, 59, 60], devoted books[61], open-source software [62, 63], mobile applications [64, 65], speaker and recognition competitions [66, 35]. The National Institute of Standards and Technology (NIST) conducted a challenge of Audio-Visual speaker recognition in 2019 (Audio-visual SRE19) [35]. The challenge provided baseline face

recognition and speaker recognition and accepted two evaluation tracks, audio-only and audio-visual, along with visual-only as an optional track. This competition submissions have indicated interesting results and started a new direction in audio-visual biometrics in ongoing NIST SRE challenges. Further, there is an ongoing multimodal biometric project called RESPECT [67] which is on the verge of producing a robust audio-visual biometric recognition system. As an application, there are smartphones for performing financial transactions (e. g. banking transactions, Google pay, e-government, e-commerce), border control [36] where AV biometrics can be deployed because they provide an ideal choice for subdued and low-cost automatic recognition. Although there are no commercial biometric systems that use only audio-visual person authentication, there are domains where multimodal biometrics are used. The dependency of the constrained environment for audio-visual data capture limits the commercial use of AV biometrics. However, looking at smartphones usage growth, which is equipped with high-quality cameras and microphones, there is a scope to use audio-visual biometrics in real-world applications.

In this survey paper, we discuss audio-visual (AV) biometrics, where speech is used along with stagnant video frames of the face or certain parts of the face [68, 69, 6, 70] or video frames of the face or mouth region (visual speech) [71, 72, 73, 74, 75] in order to improve the performance. The face and speech traits are fused either at the feature level (i,e., features are fused and fed to the classifier) or at the score level (i,e., an individual recognition system is built for each trait, and scores from the system are fused). We have discussed different types of fusion schemes used in AV biometrics. The main goal of audio-visual biometrics is to improve the robustness of recognition towards unconstrained conditions and vulnerabilities. Biometric attributes (face and speech) are prone to presentation attacks where a unimodal system produces dubious recognition results. This paper also presents several presentation attack detection (PAD) algorithms that used complimentary audio-visual information (over a single cue) to obtain robust biometric systems [76, 77, 2, 78, 79].

Few survey papers are available in the literature to provide a concise review of audio-visual biometrics, including feature extraction, speaker recognition process, fusion methods, and AV databases. Deravi [80] has reviewed the audio-visual biometric systems in application to access control. Aleksic *et al.* [51] presented a survey of audio-visual biometric methods, fusion approaches, and databases until 2006. Li has performed a survey on authentication methods based on audio-visual biometrics [52] with brief reviews and presented a comparison of audio-visual biometrics until 2012. The existing papers have also discussed some of the audio-visual biometric systems [69, 81, 1, 82, 6, 83] that are vulnerable to replay attacks.

There are survey papers only on the fusion approaches used in AV biometric data fusion [84, 53]. This survey paper presents a thorough review of all spearhead efforts in AV biometrics and presentation attack detection (PAD) algorithms.

By considering the above survey papers and emerged technologies in AV biometrics, this work contributes to the following:

1. A complete up to date review of existing AV biometric systems and detailed discussion on audio-visual databases.
2. A detailed description of different audio and visual features, fusion approaches, and achieved performances are presented.
3. A thorough review of existing presentation attack detection (PAD) algorithms for audio-visual biometrics is performed.
4. Challenges and drawbacks, emerging problems, and privacy-preserving techniques in audio-visual biometrics are presented.

The rest of the paper is organized as follows: Section 6.3 presents the general concepts of AV biometric recognition system, and section 6.4 presents the features in AV biometrics. In section 6.5, we present different approaches used in audio-visual fusion and classification. Section 6.6 discusses the existing audio-visual databases and the comparison of benchmark AV algorithms on each database. Further, section 6.7 describes PAD algorithms on AV based biometrics. Section 6.8 presents challenges and open questions in this research domain and we conclude the report in section 6.9 along with discussion of future works in this direction.

6.3 General concepts of AV biometric verification system

This section discusses different types of audio-visual biometric systems and ISO Standard (ISO/IEC JTC1 SC37 Biometrics 2016)[85] biometric components. An AV biometric recognition can be classified into two types: identification and verification. Identification is a process of finding out an individual's identity by comparing the biometric sample collected from the subject with all the individuals from the database. Verification is a process where the claimed identity is checked against a single model where the biometric sample collected from an individual is compared with the same individual's sample from the database. The AV biometric system can also be divided into two types based on audio and visual data captured. Depending on the text uttered by the speaker, the AV biometric system can be called a Text-dependent or a Text-independent system. If the AV biometric system uses static visual information (e.g., an image of a face or static faces from

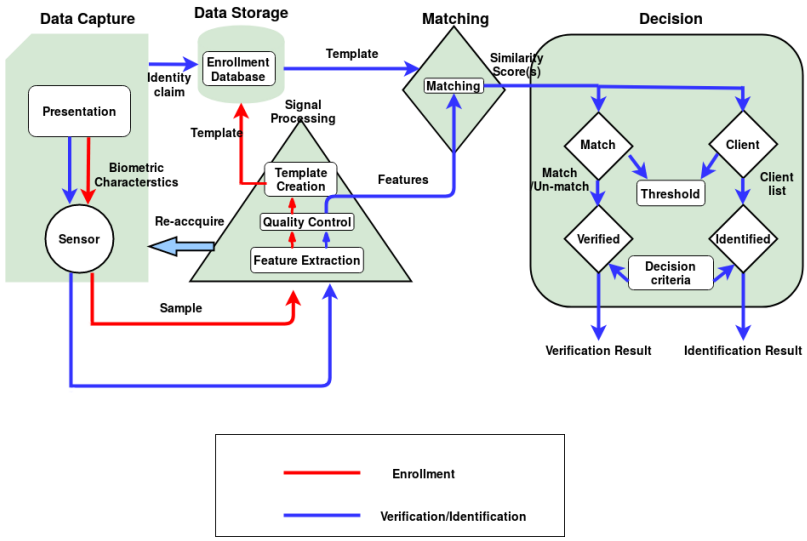


Figure 6.1: Conceptual Biometric Model inspired from ISO/IEC JTC1 SC37.

video frames) is called Audio-Visual-Static biometric systems. In contrast, AV systems using visual features containing temporal information from video frames are called Audio-Visual-Dynamic biometric systems.

6.3.1 Biometric system components

Figure 6.1 shows a block diagram of ISO/IEC JTC1 SC37 biometrics recognition diagram [85] describing two main phases of a biometrics system namely, enrollment phase (red-colored lines) and verification or identification phase (blue colored line). There are five major stages in this system: data capture, Signal processing, Data storage, Matching, and Decision making, as indicated in [85]. Data capture, signal processing, and Data Storage are used only in enrollment, and the rest of the blocks are used in both enrollment and recognition phases. The first stage is the data capture, where audio-visual biometric is captured using a sensor and the second stage is the signal processing block, which includes multiple steps. For example, segmentation and feature extraction are carried out in this step by cropping out the biometric region and extracting optimal features.

Pre-processing is a part of the signal processing block where the biometric sample is prepared for feature extraction. Pre-processing of an audio signal includes signal denoising [81], channel noise removal, smoothing [86], signal enhancement, silence detection and removal. Pre-processing a video signal consists of steps like detecting and tracking the face or any other important face regions. After feature extraction, the next sub-block in the signal processing stage is the biometric

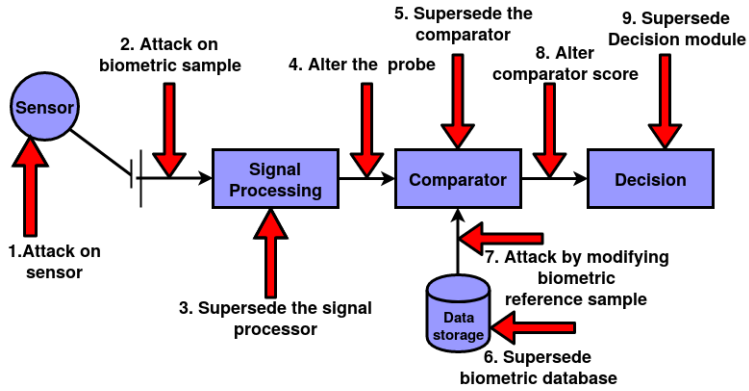


Figure 6.2: Vulnerability of AV biometric system (motivated by figure ISO/IEC 30107-1).

sample's quality control. A biometric sample is of acceptable quality if it is suitable for person recognition. According to ISO/IEC 29794-1 standardization [87], we have established three components of a biometric sample, namely *Character:* implies the source's built-in discriminative capability, *Fidelity:* the degree of resemblance between a sample and the source, and *Utility:* the samples' impact on biometric systems' all-around performance.

The next sub-block is the Data Storage, where a biometric template is created. A biometric template is a digital footnote of the peculiar characteristics of a biometric sample. Created templates are stored in the database and are used at the time of authentication. Once the sample biometric digital reference is stored in the database in the enrollment phase, the digital footnote is matched with the person seeking authentication or identification, and a binary decision, accept or reject, is made based upon a threshold both in the identification and in verification.

6.3.2 Presentation Attack Detection (PAD)

A biometric recognition system is prone to multiple types of threats. Among these, presentation attacks are considered to be one of the significant vulnerabilities. Figure 6.2 shows the generic block diagram of the biometric recognition system (in our case, audio-visual) with nine contrasting vulnerabilities, as illustrated in ISO/IEC 30107-1 [5]. The first vulnerability is at the sensor, where a pre-recorded audio or face image artifact of a lawful client is presented as an input to the sensor. An artifact as defined in ISO/IEC JTC1 SC37 Biometrics 2016 [5] is a morphed object or depiction presenting a copy of biometric characteristics or fabricated biometric patterns. This kind of attack is also known as a presentation attack.

Presentation attacks are defined as the presentation to a biometric capture subsystem with the goal of interfering with the operation of the biometric system [5]. Presentation Attack Instrument (PAI) is the biometric characteristic or the object used in a presentation attack. Presentation attacks can be divided into two types: an active impostor presentation attack and a concealer presentation attack. The active impostor attacks are a type of attack in which the attacker tends to be recognized as a different subject. This type is again divided into two types. The first type is that the intention is to get recognized as a subject known to the AV biometric system. The second type is to get recognized as the unknown person to the AV biometric system. A concealer presentation attack is the type of attack where the subject tries to avoid getting recognized as a subject in the system.

The popular presentation attacks in audio-visual biometrics are replay attacks and forgery attacks. A replay attack is performed by replaying the audio-visual recording sample in front of a biometric sensor. This can be performed either on individual modality (face video replay or audio recording replay) or both modalities at once (audio-visual replay). Forgery attack is carried out by altering the audio-visual sample to make it look like a bona fide sample of the target speaker. Audio-visual forgery includes two modal transformations. Speaker transformation, also known as voice transformation, voice conversion, or speaker forgery, is a technique for altering an impostor's utterance to make it sound like the target speaker (client). In the visual domain, face transformation aims at creating an animated face synthetically from a still image of the target client.

Presentation Attack Detection (PAD) is a framework by which presentation attacks can be identified to be classified, particularised, and communicated for decision-making and performance analysis. In literature, PAD is also termed as anti-spoofing techniques in the development of countermeasures to the biometric spoofs. In most of the existing AV biometrics literature, PAD is referred to as liveness detection; however, liveness detection is defined as the measurement and analysis of involuntary or voluntary reactions in order to detect and verify whether or not a biometric modality presented is alive from a subject at the time of capture [5]. So, from the standardization, we can infer that liveness detection is considered a subset of PAD but not as a synonym for itself.

6.3.3 Performance Metrics

In this section, we discuss the performance metrics used in the field of audio-visual biometric methods.

False Match Rate (FMR) is the percentage of impostors samples accepted by the biometric algorithm, and False Non-Match Rate (FNMR) is the percentage of bona fide samples rejected by the algorithm [85]. At a biometric system-level perform-

ance, Fale Acceptance Rate (FAR) and False Rejection Rate (FRR) are reported in the place of FMR and FNMR, respectively. Many research works used an equal error rate (EER) to represent FMR and FNMR metrics in a single value. EER is a single value at which FMR and FNMR are equal. Similarly, the Total Error Rate (TER) is the sum of FAR and FRR, and half TER (HTER) is the average of the FAR and FRR. Some algorithms mentioned the accuracy rate or error rate, which is the percentage of samples being correctly classified or incorrectly classified, respectively.

6.4 AV based Feature Extraction

This section presents a brief overview of the AV features widely employed in designing the multimodal biometric system based on face and voice. Features are the distinct properties of the input signal that helps in making a distinction between biometric samples. Feature extraction can be defined as transforming the input signal into a limited set of values. Further, feature extraction is useful to discard extraneous information without losing relevant information. The majority of the literature has treated AV biometrics as two unimodal biometrics based on visual (or face) and audio (or voice) biometric characteristics. Thus, the feature extraction techniques are carried out independently on audio and visual biometrics that are briefly discussed below.

6.4.1 Audio Features

The Audio features used in audio-visual biometric methods are classified into four categories, as depicted in Figure 6.3. The details of various types of audio features are briefly discussed in the following subsections.

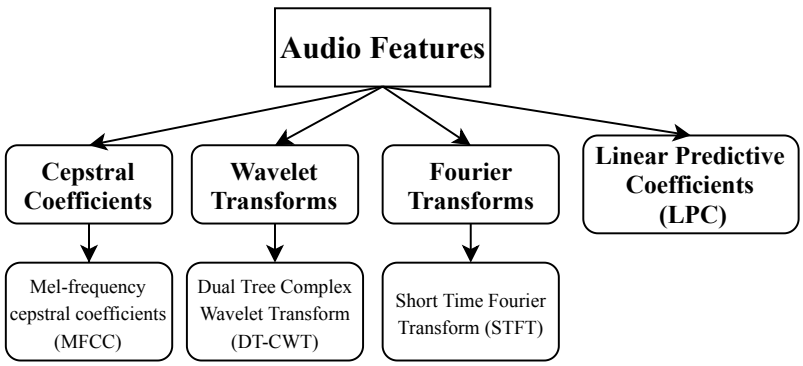


Figure 6.3: Different types of audio features used for audio-visual biometric recognition.

Cepstral coefficients

The cepstrum of a signal is obtained by applying an inverse Fourier transform of the logarithm. The logarithm is calculated from the magnitude of the Fourier transform. The advantages of cepstrum include its robustness and separation of excitation source and vocal tract system features. Robert *et al.* [34] from dialog communication systems developed a multimodal identification system where speech utterance is divided into several overlapping frames, and cepstral coefficients are extracted from these frames and used as features for AV biometric recognition.

Among cepstral coefficients, Mel-frequency cepstral coefficients (MFCC) representation is an efficient speech feature based on human auditory perceptions. MFCCs include series of operations, namely pre-emphasis (increasing magnitude of higher frequencies), framing (speech signal is divided into chunks by a window), applying Fast Fourier Transform (FFT), Mel-filtering, followed by applying DCT on log filter banks (where lower-order coefficients represent vocal tract information) to obtain the MFCCs. Mel-frequency banks approximate the human ear response more accurately than any other system, and MFCCs suppress the minor spectral variations in higher frequency bands.

MFCCs have been widely used AV for person recognition [88, 89, 1, 22, 90, 91, 81, 92, 83, 93, 94, 95, 96, 97]. Classification methods based on Gaussian Mixture Models (GMMs), Vector Quantization (VQ) have displayed a consistent speaker recognition performance using MFCCs. Experiments conducted on XM2VTS database [10], AMP/CMU database [98], VidTIMIT [99], [57] have displayed robustness of MFCCs in accurate person identification. Mobile applications [100], [101] have also used MFCCs as feature vectors. Neural network based methods [102] have examined cepstral coefficients namely i) Real Cepstral Coefficients (RCCs), ii) Linear Prediction Cepstral Coefficients (LPCC), iii) MFCCs, iv) ΔMFCCs, and v) $\Delta\Delta$MFCCs. It is observed that ΔMFCCs have performed better than others. Alam *et al.* [103],[104] have explored the usage of MFCCs in deep neural network based methods. Further, MFCCs are also used in creating i-vectors, which performed better with Linear Discriminant Analysis (LDA) and Within Class Covariance Normalisation (WCCN) [105]. In the recent works, MFCCs are used as a potential complementing feature in multimodal biometrics [106], [107].

Wavelet Transforms

The popular wavelet transform approach used in speaker recognition methods is the Dual tree complex wavelet transform (DTCWT). DTCWT uses two discrete wavelet transforms (DWT) in parallel [108], one DTCWT generates a real part of the signal other DTCWT generates the imaginary part. DTCWT is highly direc-

tional, shift-invariant, offers perfect reconstruction, and computationally efficient. Another variant of the wavelet transform is the Dual-Tree Complex Wavelet Packet Transform (DT-CWPT) [109]. It is observed that using DT-CWPT has increased the speaker identification rates in both unimodal and multimodal systems when compared to MFCCs based methods.

Fourier Transforms

The Short-time Fourier transform (STFT) is a popular Fourier transform approach used in the processing of voice as biometric data. The voice signal is a quasi-stationary signal; therefore, STFT yields better representation over a Fourier transform. In STFT, the speech utterance is segmented into frames of a smaller duration, approximately 20-30ms, and Hamming or Hanning window is superimposed on these frames before computing the Fourier transforms. The window slides throughout the signal with an overlap between the frames. Dieckmann *et al.* [75] presented a Synergetic Computer-based biometric identification system where a Hanning window covers the input signal, and STFT is applied. A power function is applied to emphasize the lower frequencies and to compress the higher frequencies.

Linear Prediction Coefficients(LPC)

The continuous-time speech signal is highly correlated. If we know the previous sample, it is possible to predict the next sample. The linear predictor predicts the next point as a linear combination of previous values. The transfer function of a linear prediction filter is an all-pole model. Linear Prediction Coefficients (LPC) model the human vocal tract as a source-filter model. Here the source is the train of impulses generated by the vibration of vocal folds, which acts as an excitation source. The filter represents the oral cavity, which models the vocal tract system, and the resulting speech signal is the convolution of a train of impulses and responses of the vocal tract system. LPCs are a compact representation of the vocal tract system and can be used for synthesizing the speech. LPCs are used for deriving the LP residual (equivalent to excitation source) with inverse filter (all-zero filter) formulation. LPCs are used for speaker recognition in AV biometric methods [110, 111, 112, 69] using Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) for classification.

6.4.2 Visual Features

This section presents a brief overview of the visual (or facial) features that are classified into four major types, as shown in figure 6.4.

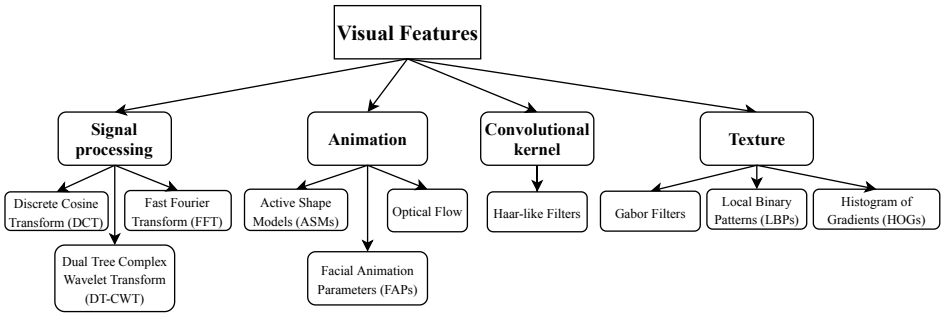


Figure 6.4: Different visual features used in audio-visual biometric recognition.

Signal Processing Based Feature Extraction

In signal processing based feature extraction, there are three different methods used in AV speaker recognition, namely Discrete cosine transform (DCT), Discrete-Time Complex Wavelet Transform (DT-CWT) iii) Fast Fourier Transform (FFT).

The discrete cosine transform (DCT) of an image represents the sum of sinusoids of varying frequencies and magnitudes. DCT has an inherent property that contains information about the image in the first few coefficients, and the rest can be discarded. DCT contains AC and DC coefficients where DC coefficients are prone to illumination changes, and hence they are discarded. However, first, few AC coefficients act as a good representation of an image. Therefore, DCTs are widely used in feature extraction and compression techniques. In the early works on audio-visual fusion for biometrics, DCTs are computed on small blocks of the image [90] and appended with the mean and variance of overlapping blocks [101]. Further, four variants of DCT methods namely DCT-delta, DCT-mod, DCT-mod-delta and DCT-mod2 [113] are examined. DCT-mod2 is formed by replacing the first three coefficients of 2D-DCT with their delta coefficients and used as feature vectors [94].

Dual tree complex wavelet transforms (DTCWT) is another feature extraction approach used for face images similar to audio features described in Section 6.4.1. DTCWT features are extracted at different depths and convolved to form a feature vector by concatenating all the rows and columns [109]. To reduce these feature vectors' dimensionality, PCA is applied, and only 24 vectors are chosen from 6 directions. Using Fast Fourier Transform (FFT), an image can be transformed into the frequency domain as a sum of complex sinusoids with varying magnitudes, frequencies, and phases. The advantage of using FFT is that the N transformed points can be expressed as a sum of $N/2$ points (divide and conquer), and thus, computations can be reused. Therefore, FFTs can be used for efficient feature extraction

methods for texture analysis. Robert W *et al.* [34] developed a novel multimodal identification system for face recognition from videos where 3D FFTs of 16 vector fields are computed with unique identifiable points from lips and faces.

Animation Based Features

The animation based visual features used in AV biometrics are active shape models (ASMs), facial animation parameters (FAPs), and optical flow features.

Active Shape Models (ASMs) are the statistical models of shape and appearance used to represent the face region in an image. Human experts annotate face images, and then a model is trained using a set of images. The ASM algorithm makes few postulates about the objects being modeled other than what it learns from the training set. ASMs not only give compact delineation of allowable variation but also avoid the unacceptable shapes being generated. ASMs are used to detect faces in images, and a neural network based AV speaker identification is employed in [102]. After successful face detection, region of interest is segmented using robust real-time skin color blob detection and radial scanline detection methods. Further, the background noise is eliminated, and finally, appearance-based face features are obtained [114]. Similarly, Bengio *et al.* [91] used point distribution models to track and extract the visual information from each image. For each image, 12 lip contour features and 12 intensity features, including their first-order derivatives, are excerpted, making a total of 48 features. Brunelli *et al.* [97] used pixel-level information from eyes, nose, and mouth regions to extract the features.

Facial animation parameters (FAPs) are a type of high-level features extracted from the lip-contour region. These high-level features have several advantages over low-level features like sensitivity to light and rotation. A 10-dimensional FAPs describing lip contours are extracted in [1], projected onto eigenspace to use in audio-visual person identification.

Optical flow is a probable motion of individual pixels on an image plane. The optical flow of the pixels can be computed by assuming Spatio-temporal variations in the image. Using a Charge Coupled Display (CCD) camera and infrared camera [75], horizontal and vertical projections of an image are computed and concatenated to a resulting gray level image and optical flow of mouth region [23]. A real-time face tracking and depth information is used to detect and recognize the face under varying pose in [88]. A dense optical flow algorithm is used to calculate the velocity of moving pixels and edges for AV person authentication in [22].

Convolution Kernel based features

Well-known object convolution kernel methods are Haar-like filters that can detect edges and lines in an image effectively. Voila-Jones face detection algorithm [20]

Table 6.1: Different audio and visual features used in AV biometric methods.

Types of visual features	Types of audio features			
	Cepstral Coefficients (MFCC)	Wavelet Transforms (DTCWT)	Fourier Transforms (STFT)	Linear Prediction Coefficients (LPC)
Signal Processing	[90], [101], [94]	[109]	-	[34]
Animation	[102], [88], [22], [97]	-	[75], [91]	[1]
Convolutional Kernel	[89], [92], [81], [115], [95], [105]	-	-	-
Texture	[93], [96] , [103], [104], [100], [107]	-	-	[110], [111],[112], [69]

used Haar wavelets to detect the most relevant features from a face such as eyes, nose, lips, and forehead. Therefore, Haar-like features are extended for the application of face recognition [89]. Visual speech features are derived from the mouth region by cascaded algorithm portrayed in [116]. Similarly, the Viola-Jones algorithm is also used for successful face recognition [92]. Asymboost is an another efficient face detection algorithm that uses a multi-layer cascade classifier to detect the face in multiple poses [117]. Under different illuminations and non-cooperative situations like pose and occlusions, face recognition is a challenging task. Therefore, histogram equalization is performed to normalize the images after the image acquisition [81]. When the face is more occluded, Haar cascade classifiers are used for detecting the eye portion of the image. An integral image representation that reduces time complexity and uses Haar-based features to perform AV person identification in [95]. Further, K-SVD (Single Value Decomposition) algorithm is used to create a dictionary for every video sample [105] by taking advantage of high redundancy between the video frames. K-SVD is an efficient algorithm for adapting dictionaries to achieve sparse signal representations of faces detected in each frame [118].

Texture based features

There are three types of texture-based features used in AV biometric methods, namely, Gabor filters, Local Binary Patterns (LBP), and Histogram of Gradients (HOG).

A Gabor filter is a sinusoidal signal with a given frequency and orientations modulated by Gaussians [119]. Since Gabor filters have orientation characteristics, they are extensively used in texture analysis and feature extraction of face images. Initial works in AV biometrics spotted face image by using best fitting the ellipse followed by identifying eyes and mouth position by topographic grey relief [110]. After successful face recognition, Gabor filters are applied to extract the features, and complex Gabor responses from filters with six orientation and three resolutions are used as feature vectors [111]. Machine learning algorithms like Support Vector

Machines (SVM) with Elastic Graph Matching (EGM) have displayed noticeable results [112] [69]. Further, the Pyramidal Gabor-Eigenface algorithm (PGE) is used to extract the Gabor features [83] [120].

Local Binary Pattern (LBP) is a textual operator that labels the pixels in an image by considering the neighboring pixels' values and assigns a binary number. LBP for a center pixel is calculated first using the window and is binarised according to whether pixels have high value than the center pixel. LBP histogram is computed over the LBP output array. For a block, one of the $2^8 = 256$ possible patterns is possible. LBP's advantages include high discriminate power, computational simplicity, and invariant to gray-scale changes. The use of LBPs has shown a prominent advantage in face recognition approaches. LBPs features are used for face recognition using a semi-supervised discriminant analysis as an extension to linear discriminant analysis (LDA) [93]. Face regions in an image are detected by localizing lip and eye regions using Hough transforms [20] [121]. LBP features are extracted on the detected faces for multimodal authentication in [96], [115]. Deep neural network based AV recognition systems [103] employed LBPs as visual features from face images that are photometrically normalized using the Tan-Triggs algorithm [122]. In further research, a joint deep Boltzmann machine (jDBM) model that uses LBPs is introduced with an improved performance [104]. The histograms of the face and non-face region using LBP features are extracted, and a biometric classifier is implemented using pattern recognition in [100] [123].

Histogram of Gradients (HOG) is another popular texture feature descriptor used to extract robust features from images [18]. HOG features are chosen over Local Binary Patterns (LBP), Gabor filters, Scale Invariant Transform (SIFT) because of the properties like robustness to scale and rotation variance, and global features. The multimodal biometrics method used HOG via Discriminant Correlation Analysis (DCA) on mobile devices [107].

Table 6.1 shows how different audio-visual features are discussed in this survey.

6.5 AV based fusion and classification

Information fusion is used to assimilate two complementary modalities with an eventual objective of attaining the best classification results. The audio-visual biometric methods have utilized many fusion approaches to complement audio and video characteristics to one another. The Figure 6.5 shows classification of audio-visual fusion methods. Fusion methods are divided mainly into three types: Pre-mapping (early fusion), Midst-mapping (intermediate fusion), and Post-mapping (late fusion). In this section, different audio-visual biometric methods are described with their corresponding performances. The methods described here in-

clude the performance of recognition without presentation attacks, i.e., the impostors are zero-effort impostors. The presentation attack detection algorithms used in audio-visual biometrics are discussed in Section 6.7.

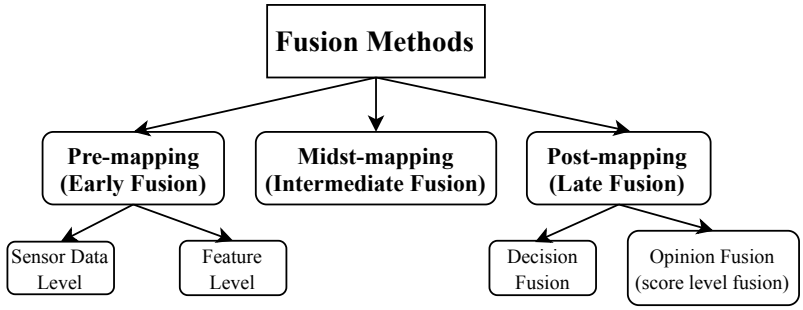


Figure 6.5: Audio-Visual fusion methods inspired from [6].

6.5.1 Pre-mapping or Early fusion

In the pre-mapping or early fusion approach, individual features from voice and face are fused to make a single set of features.

The earliest methods to use the pre-mapping of AV biometrics used fusion of static and dynamic features and used classifiers like a synergistic computer with MELT [1] algorithm [75]. The concept of synergistic computer SESAM [2] utilizes the combination of static and dynamic biometric characteristics, thus making the recognition system robust to imposters and criminal attacks. Further works explored Hidden Markov Models (HMM), which are trained using fused audio and visual features [1]. Gaussian mixture models (GMM) are also used as classifiers because of their low memory and well suitability for text-dependent and text-independent applications. GMM based classifications on concatenated features of audio and visual domains have displayed better performance than the score-level fusion [22] [90]. Some early fusion methods used clustering algorithms and PCA to reduce the dimensionality of features for efficient fusion [109]. As the cluster size is increased from 32 to 64, a higher identification rate is observed.

Laplacian projection matrix is another effective way of representing audio, and video features used in early fusion technique [92]. Laplacian Eigenmap [124] is an efficient nonlinear approach that can preserve inherent geometric data and local structure. The Laplacian matrices from both traits are fused linearly to form a single vector for audio-visual person recognition. Experiments conducted with pose estimation show an error rate of 35%. Without pose estimation, the error

[1] MELT: the prototypes of one class are *melted* into one prototype
[2] Synergetische Erkennungmittels Standbild, Akustik und Motorik

rate was 50%. The Laplacian Eigenmap fusion method outperforms the low-level fusion latent semantic analysis.

Multi-view Semi-Supervised Discriminant Analysis (MSDA) is an extension to Semi-supervised Discriminant Analysis (SDA) for feature level fusion [93]. The MSDA is inspired by a multi-view semi-supervised learning method called co-training [125]. A GMM mean adapted super vector and an LBP super vector is fused and fed into MSDA, PCA, Locality preserving projection (LPP), Linear discrimination analysis (LDA), and SDA individually. However, MSDA outperforms all other techniques because of local adjacency constraints, which can be effectively learned in different views using the same data. The synchronous measurement between audio and visual domains is examined in other works [83]. Synchronized feature vectors of size 21 are concatenated and fed to Probabilistic Neural Network (PNN). Experiments are performed at different resolutions of the face and different audio lengths (25s, 20s for training and 12.5s and 10s for testing). It is observed that the PNN method overcomes the difficulty of different frame rates for audio and visual signals and also the curse of dimensionality.

The time series vectors of face and speech provide unique characteristics of a person. The distance between data with different vector lengths is obtained using Dynamic Time Warping (DTW) [96] using the time series information from a video. The similarities between voice and face features are calculated using DTW, and multiple classifiers are fed with the similarity measures. Experiments show an authentication error of 0% for different kinds of clients. Feature-level fusion methods employed Quadratic Discriminant Analysis (QDA) for minimizing the misclassification rate, and an EER of 0.5% is obtained with the least memory and time consumption [100]. A Joint Deep Boltzmann Machine (jDBM) with a pre-training strategy and a joint restricted Boltzmann machine (jRBM) are used to model speech and face separately [104]. Then fused features were evaluated with JPEG compression and babble noise to degrade the face and speech files, respectively. The jDBM method outperforms bimodal DBM in significantly degrading conditions.

Decision voting is used for 39-dimensional audio, and video features [105]. During fusion, the standard sparsity concentration index was modified because face and speech cues are two complementary modalities and a new classification rule was derived called a joint sparse classifier. The proposed classifier outperforms the sparse representation classifier, which was used for a single modality. Discriminant Co-relation Analysis (DCA) is used to perform an early fusion of MFCCs and HOG features [107]. The DCA fused feature set is given to five different classifiers, namely Support Vector Machine, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Random Forests, and K-Nearest Neighbours. SVM

achieves the lowest EER of 20.59% among all classifiers mentioned, and SVM requires $50.9818s$ for training and $0.6038s$ for testing, which is significantly less when compared to other classifiers.

6.5.2 Midst-mapping or Intermediate fusion

The midst-mapping or intermediate fusion is a relatively complicated technique compared to the early fusion technique. In this approach, several information streams are processed while mapping from the feature space to the decision space. The intermediate fusion technique exploits the temporal synchrony between the video streams (e.g., speech signal and lip movements of videos) by which the curse of dimensionality problems with feature level fusion technique can be avoided. Examples of this type of fusion are HMMs, which can handle multiple streams of data. Asynchronous HMMs are used for text-dependent multimodal authentication in [91]. Training of AHMMs was performed using the Expectation Maximisation (EM) algorithm with clean data. Experiments were conducted on AV samples with various noise levels (0dB, 5dB, 10dB), and results display promising Half Total Error Rate (HTER) compared to audio-only and face-only modalities. In the next works, Coupled Hidden Markov Models (CHMM) are used for audio recognition and Embedded Hidden Markov Models (EHMM) or Embedded Bayesian networks for face recognition [89]. Experiments were carried out on the XM2VTS database, which resulted in error rates of 0.5% and 0.3% at various Gaussian noise levels.

6.5.3 Post-mapping or Late fusion

The post-mapping or late fusion based audio-visual fusion methods perform a data fusion on the results obtained individually from the classifiers of audio and visual domains. There are multiple ways of fusing the data in the late fusion approaches. The following sections describe all the late fusion methods used in audio-visual biometrics.

The popular methodologies considered combining the scores of audio and visual biometrics modalities using traditional mathematical rules. The sum and product rule are applied on face and voice recognition modalities built individually [126, 110]. Different face and voice recognition methods are examined, and the best performance of 87.5% subject acceptance rate was achieved by sum rule. Similarly, in [97], authors applied a weighted product approach for fusing scores from three visual classifiers and one acoustic classifier yielding an identification result of 98%. The Bayesian approach of decision fusion is another popular method used for the late fusion approach [111]. For speaker recognition, LPCs were used, and for face recognition, Gabor features are used. Experiments on the M2VTS database [56] displayed a success rate of 99.46% biometric authentication using the

Bayesian supervisor. In [112], the scores obtained from face and voice modalities are efficiently fused to obtain a new optimal score that can be used for biometric recognition. Two individual (HMM for speech and EGM for face) recognition systems are used for fusion. Experiments on the M2VTS database gave a false alarm rate of 0.07% and a false rejection rate of 0% using linear SVM. The further works focused on text-dependent and text-independent speaker recognition [69]. Co-variance matrix on LPC feature vectors and arithmetic-harmony sphericity [127] measures are employed. Different fusion schemes are experimented on the XM2VTS database for person identity verification and observed the SVM-polynomial and the Bayesian classifiers displayed better results than other methods [69].

Hidden Markov models (HMMs) provided higher accuracy in performing speaker verification. In combination with GMMs, HMMs are used with the expectation-maximization algorithm in [88]. For face recognition, Eigenvectors are employed along with GMMs, and Bayes net is used to combine the confidence scores and the conditional probability distribution. The verification experiment yields good results for the combination of both modalities with a 99.5% success rate and 0.3% rejection rate per image, and a 100% verification rate per session, and a 99.2% recognition rate per image with 55.5% rejection rate per clip. The proposed text-independent module is robust to noise variations, and the Bayesian fusion method is a simple system that can select the most trustworthy images and audio clips from each session based on confidence scores.

Cepstral coefficients have proved to be performing well by representing speaker characteristics in automatic speaker verification. A late fusion based method with three strategies is used with a matrix of the codebook of vector quantized cepstral coefficients as speech features, and a synergic computer [34] for face recognition [34]. Vector Quantization is an efficient method to characterize the speaker's feature space and is used as a minimum distance classifier. The advantage of the synergetic computer is that it can build its own features; data reduction capability makes it suitable for face recognition. The three different fusion strategies defined determines the security risk.

Gaussian mixture models (GMMs) are used over HMMs for audio-based speaker recognition and Haar based face recognizer using regularised LDA (RLDA) and Recursive FLD (RFLD) as classifiers [81]. During fusion, the probability scores obtained from classifiers are combined to get the final audio-visual probability. Experiments were performed on AVIRES corpus [128]. The RFLD classifier performs better for face recognition on the AVIRES corpus when compared to RLDA classifier with an error rate of less than 15%. GMM based methods are also used along with a universal background model (UBM) for speaker recognition and LBP

for face recognition [115]. The likelihood score from GMM-UBM and weighted distance metric on LBP are fused at the score-level. The fused method achieved an EER of 22.7% for males, 19.3% for females, and an average of 21.6%, which are far better than the EERs of individual cues. In [95], a novel Linear Regression Gaussian Mixture Models along with Universal Background Model (LRC-GMM-UBM) is used for speaker recognition. For complementing the voice utterance, a Linear Regression-based Classifier (LRC) is used for face recognition. The scores from the two classifiers are normalized and fused using the sum rule. Experiments on AusTalk database [58] give an identification accuracy close to 100% and outperforms the fusion method as shown in [129]. In another kind of late fusion approach using the same recognition algorithms [129], a combination of a ranked list, which is a subtype of decision level fusion, is used.

Session variability modeling techniques built on the GMM baseline are examined in late fusion approach [94]. Inter Session Variability (ISV) [130], Joint Factor Analysis (JFA) and Total Variability (TV) [30] are the modelling methods used in this direction. DCT coefficients are used for face recognition to model Gaussian Mixture Models (GMM) and a pre-trained Universal Background Model (UBM). Session compensation techniques include Linear discriminant analysis [131] and WCCN normalisation [132]. After session compensation cosine similarity scoring [30] and Probabilistic Linear Discriminant Analysis (PLDA) [133] are used as scoring techniques. For fusing the face and voice modalities, Linear Logistic Regression (LLR) technique is used by combining the set of classifiers using the sum rule. Experiments were performed on the MOBIO database [134] for different protocols, and results indicate that ISV performs better compared to other compensation methods and the sum rule based fusion approach of all classifiers (GMM, ISV, TV) outperforms the ISV method in all protocols.

GMM-UBM based approach is used for both face [135] and speech authentication [130, 101]. GMM-UBM uses MAP adaptation, which is prone to changes in session variability and fewer enrollment data. These drawbacks were overcome using ISV modeling proposed in [94]. A weighted sum approach is used to fuse the scores from face and speech modalities. Initially, equal weights are assigned to the two classifiers, and an LLR method is used to learn the weights on the development set. Experiments on the MOBIO database [134] have resulted in an EER of 2.6% for males and 9.7% on female subjects. Similarly, Discrete Hidden Markov Models (DHMM) are used for both audio and visual domains [102]. Experiments were performed on VALID [3] audio-visual database and observed that the proposed fusion method is very adaptable for audio-visual biometric recognition method and can be used effectively in various authentication applications.

[3]The VALID database: http://ee.ucd.ie/validdb/

The deep learning approaches have paced into the biometrics research domain in recent years. In the early research on audio-visual biometrics using deep neural networks, two restricted Boltzmann machines are used to perform unsupervised training using local binary patterns for face, and GMM super vector for voice [103]. A squashing function called softmax layer is added on the top of DBM-DNN before they are fine-tuned discriminatively using a small set of labeled training data. The authors do not mention the amount of labeled data used for fine-tuning. The sum rule was used to fuse the scores from the outputs of DBM-DNN for each cue. Experiments on MOBIO and VidTIMIT datasets resulted in EERs of 0.66% and 0.84%, respectively. Hu *et al.* proposed a multimodal convolutional neural network (CNN) architecture to perform an audio-visual speaker naming [136]. A learned face feature extractor and audio feature extractor are combined with a unified multimodal classifier to recognize the speaker. Experiments on audio-visual data extracted from famous TV shows display an improved accuracy of 90.5% over 80.8% from previous methods. Authors have also emphasized that even without face tracking, facial landmark localization, or subtitle/transcript, the proposed method achieved an accuracy of 82.9%. The latest method on a late fusion technique used similarity matrix of MFCC voice features and SVM face scores from personal devices [106]. The Lagrangian multiplier of SVM is used for fusing the scores, and the accuracy of 73.8% is obtained.

Quality assessment based score-level fusion was performed by Antipov *et al.* [137] for Audio-Visual speaker verification. For face recognition, four face embeddings, namely ResNet-50, PyramidNet, ArcFace-50, ArcFace-100, are aggregated using a Transformer aggregation model. For speaker recognition, six variants of X-vector based methods are fused using Cllr-logistic regression. The audio-visual speaker verification is performed by performing a score-level fusion of verification scores and quality of enrolment and test sample in face and speech modalities. Experiments were performed on different types of quality fusion methods compared to the baseline of sum-rule based fusion. Results indicate that using all quality estimates improve speaker verification performance.

An overview table 6.2 summarises the AV person biometric systems discussed in this survey paper.

6.6 Audio-Visual Biometric Databases

Widely variety of audio-visual databases were created by capturing talking persons' videos focusing on face and voice modalities. Multimodal databases include modalities like a fingerprint, face, iris, and biometric voice data. However, our study focuses on audio-visual databases, which include only face and voice modalities. This section presents a detailed study on both publicly available audio-visual

Table 6.2: Overview table showing features used, classifier fusion method, database, number of subjects, performance achieved, recognition type starting from the year 1995 to 2018. *TD: text-dependent, *TI: text-independent, *SEP: Standard Evaluation Protocol, *Dev: Development, *E: Evaluation, *F: Female, *M: Male

Authors	Features used		Classifier	AV fusion method	Database used	No. of subjects & sessions	Performance achieved	Recognition type
	Audio	Visual						
Brunelli *et al.* [97]	MFCC	Animation	VQ Comparision at pixel level	Weighted product	Self accquired with CCD camera	89 3 sessions	Recognition rate: 98%	Identification
Kittler *et al.* [110]	LPC	Gabor	HMM Gabor matching grid	Sum rule	M2VTS	37 16 users 21 imposters	Acceptance rate: 87.5%	Verification
Duc *et al.* [111]	LPC	Gabor	HMM EGM	Bayesian fusion	M2VTS	5328 Both client and imposter	Success rate: 99.46%	Verification
Dieckmann *et al.* [75]	Fourier transform	Optical flow	MELT algorithm	Sensor fusion	Self accquired with CCD camera	66 15 clients 26 common 25 test	Identification rate: 93% Verification rate: 99.8%	Identification Verification
Yacoub *et al.* [112]	LPC	Gabor filter	HMM EGM	SVM	XM2VTS	295 200 clients 25 evaluation imposters 70 test imposters	EER: 0.58%	Verification
Choudhury *et al.* [88]	MFCC	Optical flow	HMM Eigen vectors	Bayes net	Automated teller machine	26	Verification rate:99.5% Recognition rate: 99.2%	Verification Identification
Robert *et al.* [34]	Cepstral coefficients	FFT	VQ Synergetic computer	Sensor fusion	Self testing	150	FAR: < 1%	Verification
Nefian *et al.* [89]	MFCC	Haar-like	CHMM EHMM	Score level	XM2VTS	348 files (training) 320 files (testing)	EER: 0.5%	Identification
Bengio *et al.* [91]	MFCC	ASM	AHMM	Midst mapping	M2VTS	2 sessions (client model) 3 sessions (testing)	HTER: 15%	Verification
Isaac *et al.* [22]	MFCC	Optical flow	GMM	Feature level	XM2VTS	200 (training) 25 (test, evaluation imposters)	EER: Evaluation: 1% Test:2%	Verification
Shah *et al.* [90]	MFCC	DCT	GMM	Feature level	VidTIMIT	43 35 clients 8 imposters	FAR: 1% client FAR: 0% Imposters	Verification
Micheloni *et al.* [81]	MFCC	Haar-like	GMM RFLD	Score level	AVIRES	6	Classification error: 15%	Verification
Sugiarta *et al.* [109]	DTCWPT	DTCWT	PCA	Feature level	VidTIMIT	Session1,2 (training) session-3 (testing)	Identification: rate: 90% (TD) 93.7% (TI)	Identification
Jiang *et al.* [92]	MFCC	Haar-like	Laplacian Eigenmap	Feature level	Open web tv	10 8 (training) 2 (testing)	EER: 35% EER	Verification
Shen *et al.* [115]	MFCC	LBP	GMM-UBM LBPH	Score level	MOBIO	160 session-1 (enrolment) session 2-6 (testing)	EER: M: 22.7%; F: 19.3% Average: 21%	Verification

Authors	Features used		Classifier	AV fusion method	Database used	No. of subjects & sessions	Performance achieved	Recognition type
	Audio	Visual						
Chenxi *et al.* [83]	MFCC	Pyramidal Gabor filter	PNN	Feature level	Virtual subjects	40	Recognition rate: 100%	Identification
Motlicek *et al.* [101]	MFCC	DCT	GMM	LLR score level	MOBIO	Protocols: mobile-0, mobile-1, laptop-1, laptop-mobile-1	EER: Dev: M:1.2%; F:2.3% Test: M:2.6%; F:9.7%	Verification
Xuran *et al.* [93]	MFCC	LBP	MSDA	Feature level	MOBIO	session-1,2,3 training session-8,9,10 testing	Recognition rates: session-1: 90.6% session-2: 96.7% session-3: 97.4%	Verification
Alam *et al.* [129]	MFCC	Haar-like	LRC-GMM-UBM LRC	Ranked list	Austalk	88	Identification accuracy: 31.3%	Identification
Khoury *et al.* [94]	MFCC	DCT-mod2	GMM+ISV+TV	LLR score level	MOBIO	Protocols: mobile-0 (m0), mobile-1 (m1), laptop-1 (l1), laptop-mobile-1 (lm1)	EER: (Dev; Eval) m0: F: (1.43%; 6.30%) M: (0.92%; 1.89) m1: F: (1.64%; 6.32%) M: (0.75%; 2.06%) l1: F: (2.91%; 6.83%) M: (1.82%; 3.37%) lm1: F: (1.11%;6.32%) M: (0.64%;1.77%)	Verification
Alam *et al.* [95]	MFCC	Haar-like	LRC-GMM-UBM LRC	Sum rule	Austalk	88	Accuracy: 100%	Identification
Tresadren *et al.* [100]	MFCC	LBP	Boosted slice classiifier	QDA	MoBio	-	EER: 0.5%	Verification
Shi *et al.* [96]	MFCC	LBP	DTW LBPH	Sum rule	Self accquired	11	Authentication Error: 0%	Verification
Islam *et al.* [102]	MFCC	ASM	HMM	BPN	Self accquired	11	Authentication Error: 0%	Verification
Alam *et al.* [103]	MFCC	LBP	DBM-DNN	Feature level	VidTIMIT MOBIO	Protocol 1: train:session-1+2 test: session-3 Protocol:2 train: session-1+3 test:session-2 SEP	EER: Protocol1: 0.66% Protocol: 0.84%	Verification Identification
Primorac *et al.* [105]	MFCC	Haar-like	Joint sparse classifier	Feature level	MOBIO	SEP	Recognition rate: 0.942	Verification
Alam *et al.* [104]	MFCC	LBP	jDBM	Feature level	MOBIO	SEP	Identification: rate: 99.70 (TD) 96.30 (TI)	Identification
Memon *et al.* [106]	MFCC	SVM scores	Similarity matrix	Feature level	Self accquired	15	Accuracy :73.8%	Verification
Gofman *et al.* [107]	MFCC	HOG	SVM	Feature level	CSUF-SG5	27	EER: 20.59%	Verification
Antipov *et al.* [137]	Four CNN methods	Six X-vector variants	Logistic Regression	Score level	NIST SRE19	M/F: 15/32 (Dev) M/F: 47/102 (Test)	EER: Dev: 2.78% Test: 0.6%	Verification

biometric databases. A comparison of databases with each other can be found in Table 6.4. Some databases, like DAVID [138] is mentioned in other works, but no published work is found. Other databases like DaFEx [139] contain audio-visual data but not recorded for the application of biometrics. Therefore, they are not discussed in this report.

AMP/CMU dataset: The advance multimedia processing (AMP) lab of Carnegie Melon University (CMU) has created an audio-visual speech dataset that contains

ten subjects (seven male, three female) [4]. Each subject speaks 78 isolated words, and a digital camcorder with a tie-clip microphone is used to record [7]. The sound file and extracted lip parameters are available to the public domain, and video data is available upon request.

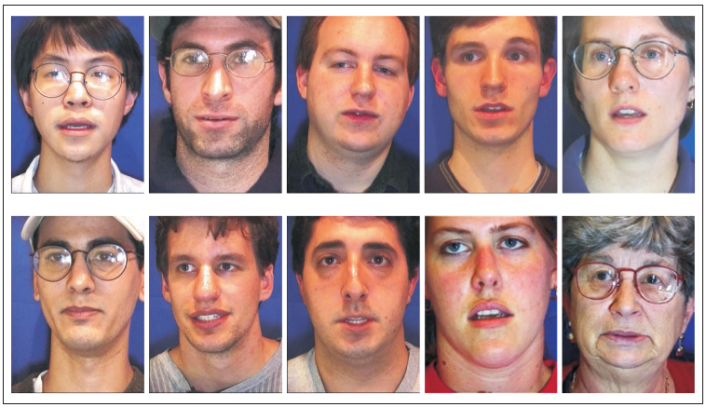


Figure 6.6: Example AMP/CMU dataset images [7].

Aleksic *et al.* [1] used 13 MFCC coefficients with first and second-order derivatives, audio features, and a visual shape-based feature vector of ten Facial animation parameters (FAPs) to develop an AV speaker recognition system. Fusion integration approach is employed with single-stream HMMs and speaker verification and identification experiments performed on the AMP/CMU dataset. The results of audio-only and audio-visual speaker recognition at different signal-to-noise ratios (SNRs) are presented in Table 6.3.

Table 6.3: Comparison of audio-only (AU) and audio-visual (AV) speaker recognition performance proposed in [1].

SNR	Identification Error (%)		Verification Error (%)	
	AU	AV	AU	AV
30	5.13	5.13	2.56	1.71
20	19.51	7.69	3.99	2.28
10	38.03	10.26	4.99	2.71
0	53.10	12.82	8.26	3.13

The BANCA database: Biometrics Access Control for Networked and E-Commerce Applications (BANCA)[5] [8] is one of the earliest audio-visual datasets used for

[4]The AMP/CMU dataset: http://amp.ece.cmu.edu/
[5]The BANCA database: http://www.ee.surrey.ac.uk/CVSSP/banca/

E-Commerce applications. Two modalities of face and voice were captured in four European languages with both high and low-quality microphones and cameras. Throughout capturing, three different scenarios, controlled, degraded, and adverse, are included in 12 different sessions for three months. The total number of subjects was 208, with an equal number of men and women. Figure 6.7 shows the example images of database captured in three different scenarios. The database is benchmarked with a weighted sum rule score-level fusion technique. The features used are DCT-mod2 for face and MFCCs for voice. The GMM models are used to perform face and voice classification, and audio-visual speaker verification obtained an equal error rate of 3.47% without impostors.

Figure 6.7: Example BANCA database images Up: Controlled, Middle: Degraded and Down: Adverse scenarios [8].

The VALID database: The aim of the VALID database is to provide robust audio, face, and multimodal person recognition systems. Therefore, the VALID database was acquired in a realistic audio-visual noisy office scenario with no control over lights or acoustics. This database is captured in five sessions with 106 subjects for a period of one month. The performance degradation of the uncontrolled VALID database is observed in comparison to that of the controlled XM2VTS database [9]. The VALID database is publicly available to the research community through the websites [6]. Figure 6.8 shows the example images from the VALID database captured in five different sessions.

[6]The VALID database: http://ee.ucd.ie/validdb/

The audio-visual experiments are performed on the VALID database to address noise problems in a single modal speaker identification [102]. A new score fusion approach is proposed using a back-propagation learning feed-forward neural network (BPN). The verification results from appearance-shape based facial features and MFCC based audio features are combined using BPN score fusion, and a speaker identification of 98.67% is achieved at SNR of 30dB.

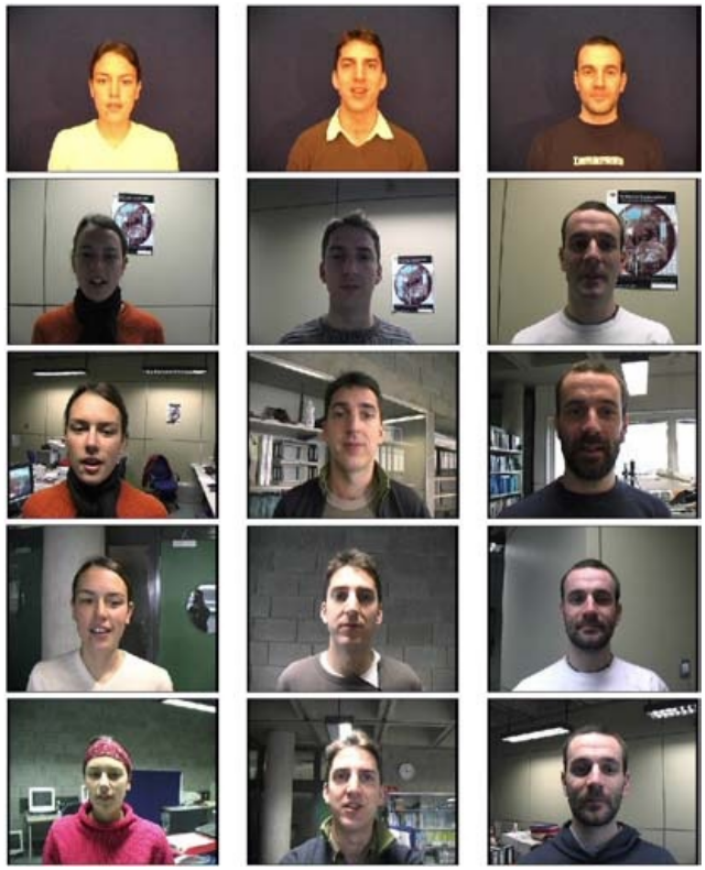


Figure 6.8: Three VALID database subject images from each of the five sessions [9].

The M2VTS database: The MultiModal Verification for Teleservices and Security applications database has developed with the primary goal of issuing access to secure regions using audio-visual person verification [140]. Five shots were taken for each of the 37 subjects, with an interval of one week between each shot. The camera used for shooting the face images is a Hi8 video camera. D1 digital recorder is utilized for recording and editing the voice. The voice recordings are captured by speakers uttering the numbers from 0 to 9 in their native language

(mostly French). The M2VTS database is available to any non-commercial user on request to the European Language Resource Agency.

Multimodal data fusion using support vector machines (SVM) method used the M2VTS database to perform audio-visual person identification [112]. The experiments display a dominance of SVM performance over Bayesian conciliation, speech only, and face only experts. This approach's face features are based on Elastic Graph Matching (EGM), and speech features are Linear Predictive Coefficients-Cepstrum (LPC-C). The total error rate (TE), which is a sum of false acceptance (FA) rate and false rejection (FR) rate, is computed, and Linear-SVM gave the least TE of 0.07%.

The XM2VTS database: An extension to the M2VTS database with more subjects and latest devices, XM2VTS (extended M2VTS) is focused on a large multimodal database with high-quality samples [10]. This database contains four recordings of 295 subjects taken during four months. Each recording contains a speaking headshot and a rotating headshot. The data comprises high-quality color images, 32 kHz 16-bit sound files, video sequences, and a 3D Model. XM2VTS database is used in many research works for AV speaker verification. The database is made publicly available at cost price only [7].

Different fusion approaches are experimented on XM2VTS database for person identity verification [69]. The elastic graph matching (EGM) based face features are computed, and two voice features, namely sphericity, and hidden Markov models (HMM), are used for six different fusion classifiers (SVM-polynomial, SVM-Gaussian, C4.5, Multilayer perceptron, Fisher linear discriminant, Bayesian classifier). It is observed that the bayesian fusion method with different combinations of face and voice features (with text-dependent and text-independent scenarios).

A coupled HMM (CHMM) is used as a classifier as audio-visual speech modeling for speaker identification. 2D discrete cosine transform (2D DCT) coefficients are used as facial features and MFCCs as acoustic features. The visual speech features are computed from the mouth region through a cascade algorithm. Finally, the audio features and visual features are combined using a CHMM. A two-stage recognition process is performed by computing face likelihood using embedded HMM and audio-visual speech likelihood using CHMM separately. The face-audio-visual speaker identification system is created by combining face and audio-visual speech likelihoods and has achieved an error rate of 0.3%.

VidTIMIT database: Video recordings of people reading sentences from Texas Instruments and Massachusetts Institute of Technology (TIMIT) corpus (VidTIMID)

[7] The XM2VTS database: http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

Figure 6.9: Front profile shots of a subject from four sessions of XM2VTS database [10].

[8] is a publicly available dataset for research purposes [57]. The dataset is captured in 3 sessions with a mean delay of 6-7 days. Each person reads ten sentences that include alpha-numerical utterances, with the first two sentences the same for all subjects. Along with the sentences, a head rotation sequence is recorded for each person in each session [99]. VidTimit dataset is used in audio-visual person recognition using deep neural network [103]. The local binary patterns (LBPs) as visual features and gaussian mixture models (GMMs) built on MFCCs have used speech features. The Deep Boltzmann Machines based deep neural network model (DBM-DNN) is used to compute scores from extracted features fused using a sum rule. The audio-visual based speaker recognition has improved the performance over the single modal recognition with an EER of 0.84%.

Liveness detection is another prominent area where VidTIMIT is employed. Gaussian mixture models [76] [2], Cross-modal fusion [141] and delay estimation methods [142] are experiments on VidTIMIT dataset to perform replay attack detection using audio-visual complimentary data.

BioSecure database: BioSecure[9] is another popular multimodal database contains different biometric modalities and can be used as a audio-visual dataset [11]. The

[8] The VidTIMTI dataset: http://conradsanderson.id.au/vidtimit/

[9] BioSecure: https://biosecure.wp.tem-tsp.eu/biosecure-database/

database consists of data from 600 subjects recorded in three different scenarios. The sample images from the database are shown in Figure 6.10.

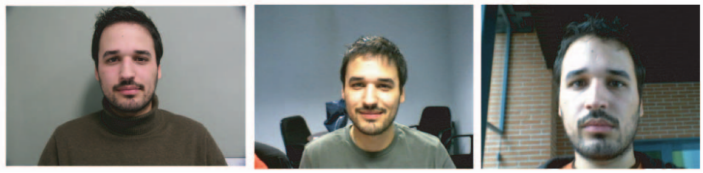


Figure 6.10: Face samples acquired in BioSecure database in three different scenarios. Left: indoor digital camera (from DS2), Middle: Webcam (from DS2), and Right: outdoor Webcam (from DS3) [11].

AVICAR [10]**:** AVICAR is a public audio-visual database captured in a car environment through multiple sensors consisting of eight microphones and four video cameras [143]. The speech data consists of isolated digits, isolated letters, phone numbers, and sentences in English with varying noise.

MOBIO database [11]**:** The MOBIO database [134] is a bi-modal (audio and video) data collected from 152 people with 100 males and 52 females. It is captured at six different sites from five different countries in 2 phases (6 sessions in each phase). This database's important feature is that it was recorded using two mobile devices: a mobile phone (NOKIA N93i) and a laptop computer (2008 MacBook).

MOBIO dataset helped in the study of person identification in a mobile phone environment. Session variability modeling is used to perform bi-modal authentication using the MOBIO database. Inter-session variance is exploited to compensate for the drawbacks of GMM-UBM based methods, and a weighted sum rule based fusion [101]. Using face features like DCT with GMM modeling and sum rule based fusion displayed an improvement in person authentication [94]. Further, deep learning methods have also used the MOBIO database for experiments on AV person recognition. DBM-DNN [103] and jDBM [104] methods are utilised on MOBIO dataset and displayed improved person identification.

MobBIO database: The MobBIO database consists of face, iris and voice biometrics from 105 volunteers (29% females and 71% males) [144]. The data capturing process took place in 2 different lights. The device used is the rear camera of the Asus Transformer Pad TF 300T for capturing 16 faces and 16 iris images. Each volunteer was asked to read 16 sentences in Portuguese for voice biometrics.

Hu *et al.* dataset: A new audio-visual dataset was recently captured by Hu *et*

[10] AVICAR database: `http://www.ifp.uiuc.edu/speech/AVICAR/`
[11] The MOBIO database: `https://www.idiap.ch/dataset/mobio`

al. [136], which is used in developing a deep learning-based feature fusion. The database is acquired from three hours of videos of nine episodes from two popular television shows with annotated subjects. Face and audio of six people from "Friends" and five from "The Big Bang Theory" are annotated and provided in this dataset. Two initial experiments are used, namely, face only recognition and identifying non-match face-audio pairs to improve audio-visual recognition performance (speaker naming). In the speaker naming process, a neural network approach is used to identify the speaker in each frame using a matched face-audio pair. This method has achieved an accuracy of 90.5%.

AusTalk database: Australian Speech Corpus (AusTalk)[12] provides the data of people reading predefined set of sentences in English [58]. The database is a part of the Big Australian Speech Corpus project consisting of speech from 1000 geographically and socially diverse speakers and recorded using a uniform and automated protocol with standardized hardware and software. A linear-regression based classifier it used for audio-visual person identification on AusTalk database achieving 100% accuracy [95].

SWAN database: The Smartphone Multimodal Biometric database was collected to meet the real-life scenarios such as mobile banking [12]. The database was captured in six different sessions and four locations using iPhone 6s and iPad Pro cameras. The database consists of audio-visual data of 150 subjects with English as a common language and Norwegian, French, and Hindi as secondary languages. Figure 6.11 shows the sample images of subjects from six sessions.

Figure 6.11: Talking face samples from SWAN database one frame from each session [12].

NIST SRE19 AV database: This database contains the videos from the VAST portion of the SRE18 development set. This database's development set is publicly available for the 2019 NIST Audio-Visual Speaker Recognition Evaluation [35]. The videos are in interview-style and are similar to the VoxCeleb database. The videos are incredibly diverse in quality and acoustics because they are recorded mostly using personal handheld devices like smartphones. The videos contain manually diarization labels as the videos may contain multiple speakers.

[12] AusTalk database: https://austalk.edu.au/

Table 6.4: Details of Audio-visual Biometric Verification Databases.

Dataset	Year	Devices	No. of subjects	Best Performing Algorithm	Accuracy
AMP/CMU [7]	2001	Digital Camcorder, tie-clip microphone	10 (7 M, 3 F)	MFCC + FAPs [1]	EER = 3.13%
BANCA [8]	2003	Webcam and Digital Camera	208 (104 M, 104 F)	DCT-mod2 + GMM [79]	EER = 3.47%
VALID [9]	2005	Canon 3CCD XM1 PAL	106 (77 M, 29 F)	BPN score fusion[102]	Accuracy = 98.67%
M2VTS [140]	2005	Hi8 camera, D1 digital recorder	37	EGM + LPC [112]	Total error rate = 0.07%
XM2VTS [10]	2005	Sony VX1000E, DHR1000UX	295	CHMM [89]	Error rate = 0.3%
VidTIMIT [57]	2009	Digital video camera	43 (24 M, 19 F)	FAPs + MFCC [1]	EER = 1.71%
BioSecure [11]	2010	Samsung Q1, Philips SP900NC HP iPAQ hx2790 Webcam, PDA	DS1: 971 DS2: 667 DS3: 713	-	-
AVICAR [143]	2010	Multiple, sensors	100 (50 M, 50 F)	-	-
MOBIO [134]	2012	Nokia N93i Mac-book	152	jDBM [104]	Accuracy = 99.7%
MobBIO [144]	2014	Asus Transformer Pad TF 300T	105	-	-
Hu *et al.*[136]	2015	-	11	Deep Multimodal Speaker Naming [136]	Accuracy = 90.5%
AusTalk [58]	2016	Black Box	88	LRC-GMM-UBM [95]	Accuracy = 100%
SWAN database [12]	2019	iPhone 6 iPad Pro	88	FaceNet+DRN [12]	EER = 3.1%
NIST SRE19 AV database [35]	2019	Multiple devices	15, 37 (M, F)	Anonymous [35]	EER = 0.44%

The 2019 NIST Audio-Visual SRE challenge has releases results of the top-performing submissions in AV recognition. The approach used by the top-performing method is unknown. However, the results show that combining face and speaker recognition systems have displayed an increase of 85% of minimum detection cost compare to face or speaker recognition system alone. The EER for AV speaker recognition achieved by the top-performing team is 0.44%.

6.7 Presentation Attack Detection (PAD) Algorithms

Audio-visual biometrics are vulnerable to various artifacts that can be generated with less cost. So it is necessary to identify and mitigate these attacks to enhance both the security and reliability of AV recognition systems. This section presents a thorough review of existing presentation attack detection (PAD) algorithms against replay attacks and forgery attacks for AV biometrics. Although there are many

attack detection algorithms in single biometrics, like ASVSpoof [40], we have only included the audio-visual PAD methods in this section. The main intention is to take advantage of bimodal biometric characteristics to optimize attack detection algorithms.

6.7.1 Audio-Visual features used for liveness detection

Many works in AV biometrics suggested the liveness detection technique, which acts as a guard for possible replay attacks against the audio-visual recognition system. The fig 6.12 shows different features used for PAD in audio-visual biometrics.

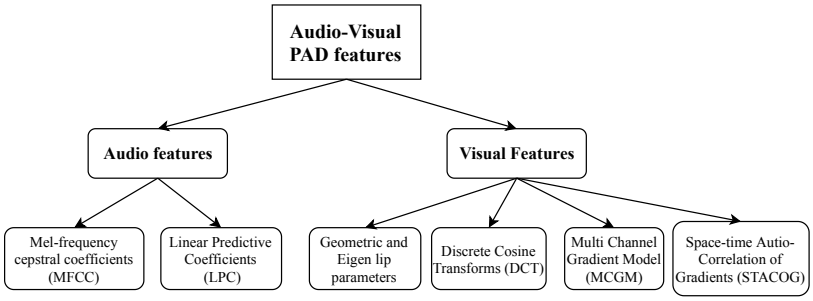


Figure 6.12: Different Audio-Visual features used in PAD.

Mel-Frequency Cepstral Coefficients

MFCCs are popular feature vectors used for both speaker recognition and liveness detection. There are different visual features used alongside MFCCs to perform reliable liveness detection. They include Geometric lip parameters and Eigen lips, Discrete cosine transforms (DCTs), Multi-Channel Gradient Model (MCGM), and Space-Time Auto Correlation of Gradients (STACOG).

Geometric lip parameters used for AV liveness verification are heights and widths of inner, outer, lower, and upper lip regions [76]. Also, Eigen lip representation is used for complementing the MFCCs parameters in this method. The advantage of this method is that the alternate color spaces in an image are exploited compared to deformed images, and it can be extended to detect and extract multiple faces and their features with different backgrounds. In further works of Chetty *et al.*, a multi-level liveness verification is proposed by exploiting correlation between the cues using MFCCs, Eigen lip, 3D shape and textures features of the face with the help of different fusion techniques [2].

DCT coefficients on lip regions are used as visual features complementing MFCCs for liveness detection in [77]. Further, a client dependent synchronous measure is introduced using the Voila-Jones algorithm [20] for detecting face region and ex-

tracting first order DCTs [145]. DCT-mod2 [113] coefficients are another face image representations computed on normalized faces for robust audio-visual biometric systems against forgery attacks [79]. Another approach used DCT coefficients extracted from mouth region with Least Residual Error Energy (LREE) algorithm [146] and MFCCs as audio features [142].

Multi-Channel Gradient Model (MCGM) algorithm is a neurological and psychological algorithm used to build artificial vision systems [147]. MCGM uses gradient methods, which computes the motion as a ratio of partial derivatives of input image brightness concerning space and time. MCGM is used in cross-modal fusion method for biometric liveness verification using kernel Canonical Correlation Analysis (kCCA) [141]. This method aims at extracting the non-linear correlation between audio-lip articulators and lip motion features from MCGM. The audio-visual cues are mutually exclusive; hence a statistical technique called Independent Component Analysis (ICA) is used. The advantages of cross-modal fusion are that it exploits mutually independent components from face and voice cues in Spatio-temporal couplings and extracts correlated information.

Space-Time Auto Correlation of Gradients (STACOG) is a motion feature extraction method that uses space-time gradients of three-dimensional moving objects in a video. STACOGs are used for measuring audio-visual synchrony to discriminate live and biometric artifact samples [78]. STACOG utilizes auto-correlation to exploit the local-relationship, such as co-occurrence among space-time gradients. STACOG also exploits local geometric characteristics and possesses shift-invariance, which is a useful property for biometric recognition.

Linear Predictor Coefficients

Linear prediction model predicts the next point as a linear combination of previous values (see section 6.4.1). In complementing LPCs as audio features, geometric lip parameters are extracted from the jumping snake algorithm [148], which achieves lip segmentation. These parameters are used with a co-inertia analysis for the liveness test in audio-visual biometrics [149]. Three video features are extracted from the width, height, and area of the mouth region. The audio-visual features used here exhibit a tight link between the lip contour and speech produced to detect liveness effectively.

6.7.2 Liveness detection methods for replay attacks

This section discusses the liveness detection methods used for identifying replay attacks. As the audio-visual biometrics contain complementary information, research works used different approaches to make use of both modalities in effectively detecting replay attacks.

Gaussian Mixture Models (GMMs) comprise audio-visual feature vectors trained for each client and used as a countermeasure for replay attacks [76]. For testing clients, a log-likelihood is computed against the client model. Three experiments were conducted for live and four replayed recordings, and performance is calculated. The results of liveness detection show that using Eigen lip projections, lip contours with MFCCs gives an EER for less than 1% for all cases.

The Co-Inertia (CoIA) and Canonical Correlation Analysis (CANCOR) are statistical methods used to measure the relationship between two multidimensional data. They are computed using Pearson correlation projections are used for liveness detection in AV biometrics [149]. Ordinary correlation analysis is dependent on the coordinate system in which the variables are described, where CANCOR and CoIA focus on finding the best coordinate system, which is optimal for correlation analysis. CoIA method has numerical stability and does not suffer from collinearity. Experiments were conducted on the XM2VTS database with two kinds of replay attacks created with the same and different sentences uttered. The result shows that CoIA method gives EER values of 14.5% and 12.5%, where CANCOR method shows 23.5% and 22.5% on replay attack 1 and replay attack 2, respectively.

Multi-level liveness verification is proposed using three different fusion techniques, namely Bi-modal feature fusion (BMF), Cross-Modal Fusion (CMF), and 3D multimodal fusion (3MF) [2]. Experiments were performed on VidTIMIT, UCBN, and AVOZES. A 10-mixture Gaussian mixture model for each client and each fusion approach is trained by constructing a gender-specific UBM and then adapting each UBM with MAP adaptation. While testing, the client's live recordings were evaluated against the client's model by calculating the log-likelihood of audio-visual vectors. Three types of replay attacks were created for testing: photo replay attack, video replay attack, and synthetic replay attack. It is observed that photo replay attacks are easy to detect compared to the other two attacks and the 3MF fusion method is more robust than other methods. The equal error rates of the three proposed techniques on three types of replay attacks is shown in Table 6.5.

Table 6.5: Performance of liveness verification techniques proposed in [2] (EER%).

Approach	Photo replay	Video replay	Synthetic replay
BMF	2.4%	6.54%	9.23%
CMF	0.29%	2.25%	3.96%
3MF	0.0155%	0.611%	1.18%

The synchronous information between audio and visual cues is an advantageous

detail that can be used in liveness detection. By measuring the asynchrony, a presentation attack can be detected. The degree of synchrony between lips and voice in a video sequence is used for liveness detection in [77]. The methods used are co-inertia analysis (CoIA) and coupled hidden Markov models (CHMM). Three different methods of CoIA were proposed, namely world training model, self-training method, and piece-wise self-training method. A CHMM is a collection of HMM which uses the Baum-Welch algorithm for training. Further, the Viterbi algorithm calculates the states for every stream and the frame likelihoods. CoIA and CHMM methods are fused using Bayesian fusion [150]. Experiments were performed on the BANCA database [8] using two protocols: controlled and pooled. Only recordings from controlled conditions are used from the BANCA database's world model in the controlled protocol. In the Pooled protocol, three conditions, such as controlled, adverse, and degraded, are used. The sum rule fusion of CoIA and CHMM methods in controlled protocol resulted in lower error rates than individual methods. The CHMM method displayed the lowest error rates in detecting replay attacks in the pooled protocol.

In similar fashion, a client dependent synchrony measure is introduced to thwart the deliberate impostor attacks [145]. For the extracted acoustic and visual features, CoIA is applied, which maximizes the covariance of AV features in the enrollment phase. While testing for the AV features, a correlation measure based on CoIA is computed. This method produces a weighted error rate (WER) of 7.7% for random impostors and 6.9% for deliberate impostors. Since WER for Random impostor attacks is on the higher side, three other fusion strategies are proposed. The first fusion strategy is the weighted sum of scores of speech, face verification, and synchrony which makes the method sensitive to deliberate impostor attacks. The second fusion strategy aims to reduce the first strategy score with a low synchrony verification score, making it robust to random impostor attacks. The third fusion strategy is an adaptive weighted sum of normalized scores. More weight is given to the synchrony verification module if the synchrony score is the least, and weight is decreased if the synchrony score is high. The three fusion strategies proposed makes the system robust to deliberate imposters.

The cross-modal fusion based on Bayesian Fusion is adapted for Liveness detection in [141]. The audio module, PCA Eigen lip module, kCCA module, and ICA module are summed up as logarithmic class conditional probability and fused under the reliability weighted summation (RWS) rule. Experiments were performed on VidTIMIT and DaFEx databases [151] using 10-mixture GMMs and log-likelihood scores. Two types of attacks were tested: static replay attack displaying the still photo and dynamic attack where faces are synthesized from still photos. The single-mode features such as MFCC, PCA, Eigen lips produce higher

errors in detecting the attacks. However, when MFCC, PCA, Eigen lips, kCCA, and ICA were fused, the method produced a promising performance.

A delay estimation method is a process of shifting audio features positively and negatively to check for the liveness in audio-visual samples [142]. Experiments are conducted on the VidTIMIT database using three types of inconsistent data and time delay based scoring. Three types of attacks include audio-video from the same subject but a different sentence, audio-video from different subject and sentence, and finally, audio-video from a different subject but the same sentence. When experimented on Co-Inertia Analysis (CoIA) and Canonical Correlation Analysis (CCA), upon adding the delay estimation method, the performance of liveness detection has improved.

The Space-Time Auto-Correlation of Gradients (STACOG) is used for measuring the audio-visual synchrony in [78]. Two cross-modality mapping approaches are used to estimate synchrony: Partial Least Square Analysis (PLS) and Canonical Correlation Analysis (CCA). PLS method [152] models the input and output onto a low-dimensional subspace. The projections are chosen such that covariance between input and output scores are maximized. CCA is a statistical method used to measure the relationship between two multidimensional data. The correlation between the audio-visual feature vectors is obtained from either CCA or PLS, as described in [145]. Experiments were performed on the BANCA and XM2VTS database with four different kinds of replay attacks. The proposed method with STACOG for visual speech features and CCA for acoustic features produced promising results proving the advantage of visual speech joint features.

6.7.3 Forgery attacks in AV Biometrics

Forgery attacks are performed by digitally transforming both voice and face cues. There are no liveness detection approaches in the literature; however, in this section, we discuss the impact of forgery attacks on AV biometrics.

The robustness of the audio-visual biometric systems against forgery attacks is examined in [79]. The forgery attacks are created using a mixture-structured bias voice transformation technique called MixTrans. This method allows the transformed signal to be estimated and reconstructed in the temporal domain. Once the transformation is defined as bias, it makes the source client vectors resemble a target client, and a maximum likelihood criterion is used to estimate the transformation parameters. Finally, a synthesis step is used to replace the source characteristics with those of the target speaker. Face transformation is performed by a MPEG-4 face animation based approach using a thin-plate spline warping. Experiments were performed on the BANCA database consisting of 7 distinct training and testing configurations. The proposed forgery attack shows an increase in EER

Table 6.6: Table showing summary of different features, methods for liveness detection, databases used and EERs achieved. (Attack type: Replay attack)

Authors	Audio-Visual features		Method used for liveness detection	Database used	EER(%)
	Audio	Visual			
Chetty *et al.* [76]	MFCC	Geometric lip parameters	Gaussian mixture model	VidTIMIT	0.6
Eveno *et al.* [149]	LPC	Geometric lip parameters	Canonical correlation analysis Co-inertial analysis	XM2VTS	12.5
Chetty *et al.* [2]	MFCC	Geometric lip parameters	GMM-UBM	VidTIMIT, UCBN, AVOZES	0.61
Rua *et al.* [77]	MFCC	Discrete cosine transform	Co-Inertia analysis Coupled Hidden Markov Models	BANCA	2.61
Bredin *et al.* [145]	MFCC	Discrete cosine transform	Co-Inertia analysis	BANCA	6.6
Chetty [141]	MFCC	Multi channel gradient model	Reliability Weighted Summation	VidTIMIT	8.6
Zhu *et al.* [142]	MFCC	Discrete cosine transform	Co-Inertia Analysis with time delay	VidTIMIT	14.6
Boutella *et al.* [78]	MFCC	Space-time auto-correlation of gradients	Canonical correlation analysis, Partial least square analysis	BANCA, XM2VTS	5.6

of the AV identity verification system when compared to the AV system with no attacks. For two groups of the dataset, EER has increased from 4.22% to 11% and from 3.47% to 16.1% with and without forgery attacks, respectively. The vulnerability of AV biometrics to forgery attacks is high, and there is a strong requirement for forgery attack detection methods in AV systems.

Table 6.6 summarises the different AV features, PAD algorithms, different databases, EERs achieved in the AV recognition system.

6.8 Challenges and Open questions

Audio-visual biometrics has gained intensive research efforts in terms of developing novel recognition systems and PAD algorithms to thwart the artifacts. Despite all these, there are several challenges and open questions in AV biometric methods. In this section, we discuss some well-known challenges and open research questions.

6.8.1 Databases and Evaluation

The publicly available databases [8, 9, 140, 10, 57, 11, 134, 144] are usually recorded using limited number of devices and sessions. The lack of variance in biomet-

ric data challenges the development of robust AV biometric algorithms. This gives rise to the problem of generalization of a biometric algorithm. For example, in smartphone biometrics, it is necessary to have an AV-based recognition algorithm that is adaptable to changes in the recording device. Therefore the databases recorded using multiple devices and in different sessions help in improving the robustness of recognition systems. Further, the mismatches arise when the enrolled sample is from one type of device and tested with another. The change in devices also introduces the problem of cross-device recognition errors. There is a requirement of AV databases that includes different types of biometric dependencies like device, lighting, background noise.

Presentation Attack Database for AV Biometrics

Available databases are also limited in terms of various kinds of presentation attacks. Moreover, data for many attacks specified in the literature are not publicly available. There are new kinds of attacks being generated that pose a huge threat to biometric systems in both audio [40] and face [38]. For example, voice impersonation has shown to be causing a considerable vulnerability to automatic speaker recognition [153]. However, the databases or protocols to create such attacks are not publicly available. Therefore, this is a hindrance for the researchers to develop robust PAD algorithms. AV biometric systems can also be attacked in either of the audio, visual, or combined audio-visual (bi-modal) domains. Depending on the authentication algorithm, a good bi-modal attack can pose a severe threat to the AV system. So there is a need to create a high-quality presentation attacks database and make it available for research.

Face Recognition under Varying Illumination Conditions

Many researchers have used the XM2VTS [10], VidTIMIT [57] databases for implementing AV recognition systems. These databases are recorded in different sessions using different devices, but they do not contain biometric data with varying illuminations. The effect of varying illumination is among several bottlenecks in face recognition research. The proposed approaches using these databases might not perform well when the lighting changes are introduced. This problem is discussed in MOBIO [134] database, which contains samples with varying illuminations. So there is a need to include this dependency in AV biometric methods with a wide variety of illumination changes.

Usage of Advanced Sensors in Data Capture

Advancements in sensor technology and computer vision have made it possible to capture images at multiple wavelengths. Most of the visual sensors capture visible wavelengths of light (e.g., RGB images). Multi-spectral sensors can capture

wavelengths that spread at different wavelengths of the spectrum. Some advanced sensors can even capture signals at near-infrared radiation, short-wave radiations, and infrared radiation. The available literature and databases for traditional face recognition methods are based on the visible spectrum and face problems like pose variations and illumination changes. It has been proved that near-infrared image capturing improves face recognition performance. Therefore, it will be beneficial to use multi-spectral sensors in AV biometrics to overcome the problems in limited visible spectrum wavelengths.

Multi-lingual Speaker Recognition

The dependency of speaker recognition on the speaker's language has been observed in the recent works [154]. The mismatch of languages of speech samples in training, enrolling, and testing is a challenging problem in AV biometrics. Therefore, a multi-lingual AV biometric system is required for active research on this problem. A multi-lingual speaker recognition system aims to recognize a person based on speech features independent of language. It is observed that there are no previous works on AV biometrics performed the task of a multi-lingual speaker recognition system. So there is a broad scope for including the language dependency experiments in AV biometrics where the problem with language can be overcome with the complementing visual part.

6.8.2 AV Biometrics in Smart devices

Smartphone usage has grown from essential communication to multipurpose usage in the past decade. The key features of recent smartphones include mobile transactions, digital ID, and sensitive multimedia data transfer. The critical information involved in smartphone functionality requires more secure access than passwords or key phrases. Biometrics have come in to play in order to provide higher security in smartphone applications. Major smartphone vendors have deployed biometric sensors and recognition systems into smartphones (e.g., Touch ID, Face ID). Due to high security and easy to use, many third-party applications use the in-built biometric technology (e.g., Apple Pay). Banking applications like the iMobile app from ICICI bank uses fingerprint or face recognition for secure login [13]. However, the variance in devices and capturing situations in mobile environments restrict advanced biometric recognition. AV biometrics can solve some of the problems faced by unimodal recognition systems and provide better security in smartphones. There are multiple challenges around efficient utilization of AV biometrics in smartphones. Smartphones comes with wide variety of cameras and microphones to record the biometric characteristics. Therefore, the recognition

[13] https://www.icicibank.com/mobile-banking/imobile.page?#toptitle

algorithm should comprehend the huge variance in capturing channels and also adaptable to new devices.

Biometrics have been used in Internet of Things (IoT) devices to provide easy and secure control of the devices. The IoT devices contain sensitive information and use biometric technologies to protect the privacy [155]. However, the biometrics sensors embedded into IoT devices come with challenges like complex architecture in IoT infrastructure. The diverse set of devices and applications in IoT requires question the security provided by biometrics based authentication. Minute vulnerability in biometrics can pose a severe threat to the critical functionality of IoT. AV biometrics can provide a solution to the problems by providing better security than unimodal methods.

6.8.3 Privacy preserving techniques in AV biometrics

Biometric characteristics are unique to the person and cannot be changed. If the biometric template is compromised, there is a breach in the privacy of the individual. So the template security is the most crucial part of the biometric system. For tackling privacy protection, template protection techniques can be classified into three main categories: i) cancelable biometrics [156] consists of intentional, repeatable distortions of the biometric signal which are irreversibly transformed, ii) cryptobiometrics [157] where a key is generated from cryptographic algorithms, iii) biometrics in the encrypted domain [158] where homomorphic encryption techniques are applied to protect the biometric data. There are ample amount of literature available for face images in cancelable biometrics and in cryptobiometrics [159] also in homomorphic encryption [160, 161, 162]. And also there are limited amount of literature available for speech in the cancelable biometrics [163, 164, 165, 166, 167] and in homomorphic encryption [168, 169]. The privacy-preserving techniques in AV biometrics have not received much attention in the audio-visual domain equivalent to other biometrics. To the best of our knowledge, the three above mentioned techniques are not addressed for AV biometrics. Hence this an open problem that can be addressed in the AV domain.

De-identification in AV Biometric System

De-identification is defined as concealing or removing the identity of the person or replacing it with a surrogate personal identity to prevent indirect or direct identification of the person. De-identification is an important tool to protect privacy, one of the most important social and political issues of today's information society. Face de-identification in still images has a lot of literature available starting from 2000 [170] and with the introduction of deep learning, GAN-based generative models have become the benchmark for learning faces. Similarly, face de-identification in videos has also gained interest [171]. The other cue speech has

also gained much interest in de-identification. Speech de-identification is mainly based on voice transformation. Voice transformation refers to modifications of the non-linguistic characteristics (voice quality, voice individuality) of a given utterance without affecting its textual content. There are lot of literature available for voice de-identification for both text-dependent [172] and text-independent case [173]. The de-identification in AV biometrics has not received much heed when compared to individual cues. Hence it is an open problem that can be addressed.

6.8.4 Deep Neural Network (DNN) based recognition

There are only two papers [103, 104] using DNN based methods for AV biometric recognition system where [103] used two different databases, and the later used only one database. The advantages of single capture bi-modal systems like AV biometrics can be exploited with deep learning approaches by developing an intelligent system using synchronous information. Using synchronous features makes it easy to avoid the problems caused by lighting, channel noise, and some presentation attacks. Therefore, there is a open research scope for utilizing advanced deep neural network based methods to develop efficient AV biometric recognition systems.

6.8.5 Performance Evaluation for AV biometrics

The AV biometric research brings two types of biometric modalities into a single system. There are performance evaluation standards for testing biometric systems by ISO/IEC [85]. The most common metrics used by AV systems are FAR, FRR and EER. However, some research works on AV biometrics used different performance evaluation methods. Similarly, there are standard metrics by NIST namely False positive identification rate (FPIR) and False Negative identification rate (FNIR) for biometric vendor technology evaluations like face (FRVT) [174], iris (IREX) [175] and fingerprint (FpVTE) [176]. It is necessary to have a common evaluation protocol used in all the works to compare and comment on various AV biometric research works. In the case of vulnerability assessment, the impostor attack presentation match rate (IAPMR) has been a standard metric from ISO/IEC [5]. However, it is observed that EER has been used in most of the liveness detection methods. Although it makes it easy to compare different works, the EER values do not explicitly show the attacks' vulnerability. Therefore, there is a strong requirement of performance evaluation protocols to test the entire biometric system and include the effect of each cue on the whole system.

6.9 Conclusions and Future works

Biometrics based person recognition has been used in multiple domains ranging from smartphone access, mobile transactions to border control checks. The vulner-

abilities in unimodal biometric systems (audio or face only) make authentication systems prone to attacks questioning biometric recognition systems' resilience. The audio-visual biometric recognition systems have evolved to overcome these problems. The AV biometric methods take advantage of the complementary information present in correlated biometric cues, face, and voice. Over the year, multiple research works focused on AV biometrics and proposed efficient person recognition approaches. This survey paper has discussed how two complementary cues can provide critical information for AV biometric person recognition system. At first, we have presented the introduction on AV biometrics, discussed how different they are from other multimodal systems, and concepts of an ISO standard biometric recognition system and presentation attacks. Later, we have classified the different types of features used in AV biometric systems in both audio and visual domains, indicating their importance, advantages, and disadvantages. We have described the different approaches of information fusion of the two modalities used in AV biometric systems. We reviewed several AV biometric recognition systems that appeared in the literature and presented their experimental results.

We then shifted our focus onto the different databases used in the AV recognition system and presented a detailed discussion of the devices and capturing methods used in each of them. A comprehensive table is presented listing the best performing algorithm on each database. Further, we presented the PAD algorithms on AV based biometric systems. We have studied different feature extraction methods used for liveness detection and discussed the performance of detecting replay attacks. AV biometrics is a hot topic of research, with many accomplishments and exciting opportunities for further research and development. Keeping this in mind, we have discussed the challenges with several open problems and mentioned possible future research directions. Overall, this article can serve as a quick reference for AV biometric recognition systems and related PAD algorithms for beginners and experts.

6.9.1 Future Works

The detailed study on AV biometrics pointed out the challenges and open problems in this field. To overcome the challenges and solve open problems, the possible future works in this direction are briefly mentioned as follows.

- A novel database of AV biometric data can be implemented, including multiple dimensions like multiple languages, sessions, devices, and presentation attacks.
- State-of-the-art algorithms can be developed for defying the dependencies and vulnerabilities in AV biometrics.

- The advantages of AV biometrics like the correlation between face and voice can be exploited exclusively to overcome the generalization problem. This leads to new paths like visual speech or talking face biometrics.
- The growth of smartphone applications for sensitive usage can make use of AV biometrics. This direction needs a research focus on implementing AV based person recognition in a mobile environment.
- The multimodal biometrics requires special attention in protecting the stored sensitive biometrics data.

Chapter 7

Article 2: Multilingual Audio-Visual Smartphone Dataset And Evaluation

Hareesh Mandalapu, Aravinda Reddy PN, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, SR Mahadeva Prasanna, and Christoph Busch. "Multilingual Audio-Visual Smartphone Dataset And Evaluation." *IEEE Access*, doi: 10.1109/ACCESS.2021.3125485, 2021.

7.1 Abstract

Smartphones have been employed with biometric-based verification systems to provide security in highly sensitive applications. Audio-visual biometrics are getting popular due to their usability, and also it will be challenging to spoof because of their multimodal nature. In this work, we present an audio-visual smartphone dataset captured in five different recent smartphones. This new dataset contains 103 subjects captured in three different sessions considering the different real-world scenarios. Three different languages are acquired in this dataset to include the problem of language dependency of the speaker recognition systems. These unique characteristics of this dataset will pave the way to implement novel state-of-the-art unimodal or audio-visual speaker recognition systems. We also report the performance of the bench-marked biometric verification systems on our dataset. The robustness of biometric algorithms is evaluated towards multiple dependencies like signal noise, device, language and presentation attacks like replay and synthesized signals with extensive experiments. The obtained results raised many concerns about the generalization properties of state-of-the-art biometrics methods

in smartphones.

7.2 Introduction

With the advances in biometrics, the usage of passwords and smart cards to gain access into several control applications have been slowly depreciated. Henceforth for reliable and secure access control, biometrics have been deployed in various applications, including smartphone unlocking, banking transactions, financial services, border control, etc. The biometrics in access control applications improve trustworthiness and enhance user proficiency by verifying who they are. A biometric system aims to recognize the person based on their physiological or behavioural characteristics based on ISO/IEC 2382-37. The physiological characteristics include the face, iris, fingerprint etc., and behavioural characteristics include speech, keystroke, gait etc.

Smartphone biometrics has grown expeditiously over the years. The number of smartphone users crossed 3 billion in 2020 and is expected to increase in millions in the coming years. According to the Mercator Advisory Group report, 66% of smartphone users are expected to use biometrics for authentication by the end of 2024. In 2020, 41% of smartphone users used biometrics which was 27% in 2019. Among different biometric modalities, fingerprint-based authentication is at the top. However, the amount of users for face and biometrics has been increasing. Voice-based recognition increased to 20% in 2020, from 11% in 2019 and face recognition jumped to 30% in 2020, from 20% in 2019. The application of smartphone biometrics has been widely used in mobile banking, e-commerce, remote identification etc.

Different types of smartphones like Android, iPhone and blackberry provide unimodal applications based on either fingerprint, iris or face recognition, and recently speech has been added as a biometric cue for authentication purposes. The built-in biometrics are not fixed for all smartphones. For example, some smartphones come with fingerprint, and some include face recognition. The captured uni-modal biometrics like face or iris comes with several problems like low quality, variations in pose, problem with illuminations, background noise, low spatial and temporal resolutions of video [21]. Therefore, this problem is addressed in multimodal biometrics by taking advantage of default sensors like cameras and microphones. Multimodal systems like audio-visual biometrics utilize the complementary information of face and speech and exploit the user-friendly capture of face and voice in a single recording. Audio-visual biometric data capture is cost-effective and can be carried out without additional sensors (e.g., fingerprint reader or iris camera).

The applications based on biometrics in smartphones has several advantages but also exist several challenges. The key challenges are the robustness and generalizability of a biometric system caused by algorithm dependencies and evolving presentation attacks. The aforementioned challenges are the main problems that circumscribe reliable and secure smartphone-based applications. The first challenge is the algorithm dependencies which limits the interoperability of a biometric algorithm across multiple types of smartphones. Interoperability is defined as the ability of a biometric system to handle variations introduced in the biometric data due to different capture devices. Due to different kinds of smartphone sensors, capturing conditions and human behaviour. The dependency of the biometric algorithm on particular data properties limits the robustness of optimal recognition. Therefore, it is very challenging to develop a conventional biometric method for a wide variety of smartphones.

The second challenge is from the presentation attacks or also called spoofing attacks and indirect attacks, which are comprehensively explained in [38] for face and in [21] for audio-visual. Presentation attacks are defined as the presentation to a biometric capture subsystem with the goal of interfering with the operation of the biometric system [5]. Presentation attacks have become easy to create and use as a concealer or impostor towards the target subject. Growing presentation attacks and limitations in smartphone sensors cause major problems questioning the performance of smartphone biometrics.

The factors above motivated research on the study of smartphone biometrics towards the key challenges. In this direction, to examine the challenges, we need a smartphone biometrics database with different attributes. There are few biometric databases have been created using smartphones in both uni-modal [177] and multimodal biometrics [134, 12]. However, the existing databases are limited with several devices, languages and sessions. Therefore, we have created a multilingual audio-visual smartphone (MAVS) dataset considering smartphone devices, sessions, speech languages and presentation attacks. The novel dataset contains audio-visual biometric data of 103 subjects (70 male, 33 female) captured in three sessions with variable noise and illumination. Each subject utters six sentences, each in three different languages and recorded in five different smartphones. We have also created two types of presentation attacks in both audio, video and audio-visual scenarios. The first type of attack is a physical access attack which is created by replaying an audio-visual sample on a display-speaker setup and recorded using a smartphone. The second attack is a synthesized attack where audio and video are created separately via speech synthesis and face-swapping.

Further, we have benchmarked the dataset by performing extensive experiments in two directions. The first direction is to observe the biometric algorithm dependen-

cies concerning device, illumination, background noise and language. The second direction is to examine the vulnerability towards presentation attacks. The baseline presentation attack detection methods in both audio and visual domains are included in this work. The biometric recognition algorithms are chosen from the state-of-the-art methods from the literature. The experimental results are presented in ISO/IEC biometric standards [85] with pictorial representations and detailed discussion.

The rest of the paper is organized as follows. Section 7.3 presents the related work in audio-visual datasets with sample images and discussion of results. The detailed description of the multilingual audio-visual smartphone (MAVS) dataset created in this research is presented in Section 7.4. Section 7.5 describes the performance evaluation protocols used in bench-marking the MAVS dataset. Section 7.6 presents the experiments performed and results obtained and Section 7.7 concludes this paper with discussion on the future work.

7.3 Related Work

The sensitivity of data in smartphone utilization has made the usage of biometrics a critical feature. Therefore, the research in smartphone biometrics has obtained much attention in recent years. The built-in biometric sensors provide the necessary authentication for many smartphones. However, the inconsistency of performance in these devices encouraged a new direction of biometric recognition using the default sensors like camera and microphone. In this direction, few audio-visual smartphone biometric datasets have been developed by capturing talking subjects' videos. Multimodal biometric databases captured modalities like a finger photo, face, iris photo, and speech data. However, considering the standard sensors in all smartphones, we studied only audio-visual databases, including face and voice. In this section, we present a comprehensive study on audio-visual biometric databases. A detailed study on all audio-visual biometric databases is performed in [21] by Mandalapu *et al.* along with a comparison of best-performing algorithms. In this section, we present some audio-visual databases in detail.

Early audio-visual biometric datasets are created by the advanced multimedia processing (AMP) lab of Carnegie Melon University (CMU) [1]. With ten subjects, each speaking 78 isolated words, the recording is taken by a digital camcorder with a tie-clip microphone [7]. The dataset is made publicly available with sound files and lip parameters. Although the number of subjects is low, this dataset assisted in developing a visual shape-based feature vector for audio-visual speaker recognition in [1]. Biometrics Access Control for Networked and E-Commerce

[1]The AMP/CMU dataset: http://amp.ece.cmu.edu/

Applications (BANCA)[2] [8] is developed for E-Commerce applications. Important features in this database are multiple European languages captured using both high and low-quality devices under three different scenarios: controlled, degraded, and adverse. Also, the total number of subjects was 208, with an equal number of men and women. Figure 6.7 shows the sample images of this database from three different scenarios.

Figure 7.1: Example BANCA database images Up: Controlled, Middle: Degraded and Down: Adverse scenarios [8].

The goal of multimodal biometrics is to improve the robustness of the recognition/verification process. The VALID database was created in a realistic audio-visual noisy office room under uncontrolled lighting and acoustic noise. The VALID database is publicly available to research purposes [3]. The MultiModal Verification for Teleservices and Security (M2VTS) applications database has been developed for granting access to secure regions using audio-visual person verification [140]. An extension to the M2VTS database is XM2VTS (extended M2VTS) with focus on high-quality biometric samples [10]. It contains high-quality face images, 32 kHz 16-bit audio files, video sequences, and a 3D Model. The database is publicly available at cost price [4].

Video recordings of people reading sentences from Texas Instruments and Mas-

[2]The BANCA database: http://www.ee.surrey.ac.uk/CVSSP/banca/
[3]The VALID database: http://ee.ucd.ie/validdb/
[4]The XM2VTS database: http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

Figure 7.2: Front profile shots of a subject from four sessions of XM2VTS database [10].

sachusetts Institute of Technology (TIMIT) corpus (VidTIMID) [5] is a publicly available dataset presented in [57]. A distinctive part of VidTIMIT dataset is that it also contains head rotation sequence for each person in each session [99]. BioSecure[6] is a popular multimodal database that also comprises of audio-visual dataset [11]. The database consists of data from 600 subjects recorded in three different scenarios. The sample images from the database are shown in Figure 6.10.

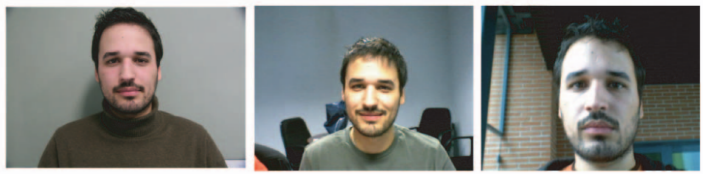


Figure 7.3: Face samples acquired in BioSecure database in three different scenarios. Left: indoor digital camera (from DS2), Middle: Webcam (from DS2), and Right: outdoor Webcam (from DS3) [11].

The aforementioned audio-visual datasets are captured with different types of sensors. In some cases, the audio and video capturing sensors are two different devices, and

[5]The VidTIMTI dataset: http://conradsanderson.id.au/vidtimit/

[6]BioSecure: https://biosecure.wp.tem-tsp.eu/biosecure-database/

the data is presented separately. However, in smartphones, the built-in camera and microphone can be used to create audio-visual data. The MOBIO database [7] [134] is a audio-visual data created using a mobile phone (NOKIA N93i) and a laptop computer (2008 MacBook). MOBIO dataset helped in the study of person identification in a mobile phone environment [101]. In a similar fashion, the MobBIO database is developed by Sequeira *et al.* in [144]. The sensors used in this work are the rear camera of the Asus Transformer Pad TF 300T.

Figure 7.4: Talking face samples from SWAN database one frame from each session [12].

The Smartphone Multimodal Biometric database was collected for the application of mobile banking [12]. The real-world scenarios are attributed in this database with multiple sessions and languages using iPhone 6s and iPad Pro. Along with audio-visual data, the SWAN database also contains face, eye region, finger photo and voice data. Presentation attacks are also provided as a part of this database. Figure 6.11 shows the sample images of subjects from six sessions.

The existing databases on audio-visual biometrics provide limited variance in addressing the problem of robustness—most databases on session variance but not on device variance and language dependency. Alongside, presentation attacks are growing widely and displaying a huge impact on the optimal performance of biometric algorithms. We have formulated advanced protocols to create a multilingual audio-visual smartphone (MAVS) database considering all these problems. In this direction, the significant contributions of this paper are mentioned as follows.

1. A novel multilingual audio-visual smartphone dataset will be made available for research purposes. The uniqueness of this dataset is described below.
    - Biometric data from 70 male and 33 female subjects from various backgrounds.
    - Three language speeches and three sessions (variable illumination and background noise) for all the subjects.
    - Data recorded on multiple smartphone devices: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8.

[7] The MOBIO database: https://www.idiap.ch/dataset/mobio

Table 7.1: Details of Audio-visual Biometric Verification Databases.

Dataset	Year	Devices	No. of subjects	Biometric	Availability
AMP/CMU [7]	2001	Digital Camcorder, tie-clip microphone	10 (7 M, 3 F)	Face, voice	Free
BANCA [8]	2003	Webcam and Digital Camera	208 (104 M, 104 F)	Face, voice	Free
VALID [9]	2005	Canon 3CCD XM1 PAL	106 (77 M, 29 F)	Face, voice	Free
M2VTS [140]	2005	Hi8 camera, D1 digital recorder	37	Face, voice	Free
XM2VTS [10]	2005	Sony VX1000E, DHR1000UX	295	Face, voice	Free
VidTIMIT [57]	2009	Digital video camera	43 (24 M, 19 F)	Face, voice	Free
BioSecure [11]	2010	Samsung Q1, Philips SP900NC HP iPAQ hx2790 Webcam, PDA	DS1: 971 DS2: 667 DS3: 713	Face, Fingerprint Voice, Signature	Paid
MOBIO [134]	2012	Nokia N93i Mac-book	152	Voice, Face periocular	Free
MobBIO [144]	2014	Asus Transformer Pad TF 300T	105	-	-
Hu *et al.*[136]	2015	-	11	Audio-Visual	Free
SWAN database [12]	2019	iPhone 6 iPad Pro	88	Face, Periocular, Multilingual Voice Presentation Attack dataset	Free
MAVS databset	2021	iPhone 6, iPhone 10, iPhone 11 Samsung S7 and Samsung S8	103 (70 M, 33 F)	Face, Multilingual Voice Presentation Attack dataset	Free

- Three unique and three common sentences for each subject, each device, each language and each session.
- Two types of presentation attacks are created, each in physical access and logical access scenarios.

2. Benchmarking the dataset with state-of-the-art face recognition, speaker recognition algorithms and score-level fusion biometric methods.

3. Evaluating the vulnerability of presentation attacks on state-of-the-art biometric verification and testing baseline presentation attack detection methods.

7.4 Multilingual Audio-Visual Smartphone (MAVS) Dataset

7.4.1 Acquisition

In data acquisition, we have used five smartphone devices, namely iPhone 11, iPhone10, iPhone 6s, Samsung S7 and Samsung S8. The data capturing is a self-assisted process where the speaker handles the mobile device and records the biometric data. For the process of data capturing, a mobile application has been used

in both iOS and Android devices. The application provides a simple interface that assists the speaker to provide audio-visual data, as shown in Figure 7.5. A pre-defined text appears on the screen for a limited time for each sample. The speaker reads the text while the data is being recorded.

Figure 7.5: Mobile application (iOS) interface for data capturing.

7.4.2 Participant details

We have obtained 70 male and 33 female participants for the data collection. The average age of the participants is 27 years. All participants are of Indian origin with medium to expert range fluency in speaking the three languages (English, Hindi and Bengali). All participants are informed about the data acquisition protocol and are instructed to use the mobile application by self-assisting the data capture. Each session, the participant is given five mobile devices, one after the other, and audio-visual data of 6 sentences in three languages is recorded.

7.4.3 Data details

Each participant records six sentences in each language. Three of the sentences are the same for all subjects, and the other three sentences have a unique part for each subject. The six sentences in the English language are mentioned below, and the blank spaces are filled with unique fake text for each subject. Similarly, translated sentences for the other two languages are presented in their corresponding script.

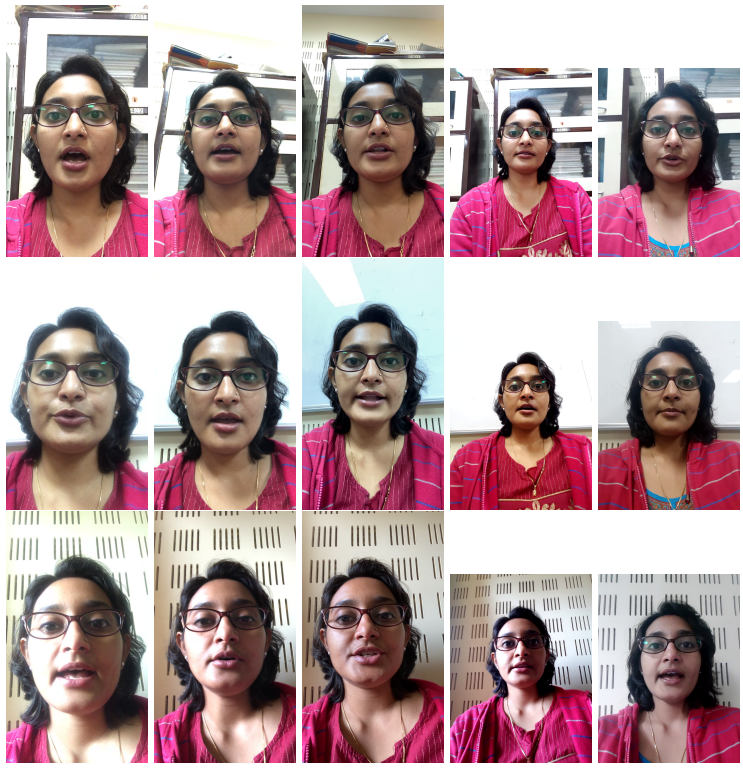


Figure 7.6: Audio-visual data samples (1 frame of a talking face). Left to Right: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8. Top row: Session 1, middle: Session2, bottom: Session3.

1. My full name is `fake name`.
2. I live at the address `fake address`.
3. I am working at IIT Kharagpur.
4. My bank account number is `fake number`.
5. The limit of my account is 10,000 rupees.
6. The code for my bank is 9876543210.

Data is captured in three sessions with three different lighting and noise environments. In session1, there is no noise, and uniform lighting is used. This data can be used as clean data for enrollment purposes. Session2 has continuous controlled noise from a portable fan intentionally put near the data capturing process

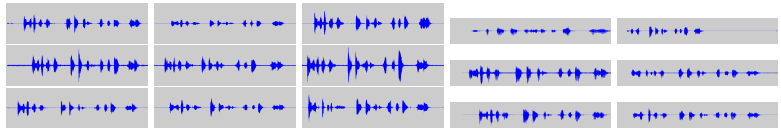


Figure 7.7: Audio data sample for speaker recognition. Left to Right: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8. Top row: Session 1, middle: Session 2, bottom: Session 3.

and different lighting than session1 but with uniform illuminance. Session3 has uncontrolled noise from natural background and nonuniform lighting where certain parts of the participant's face are dark. The order of sentences, languages, and mobile devices used during data capture is kept the same for all the sessions. The sample video data can be seen in Figure 7.6 (one frame per session, the device is presented for convenience). The waveform of audio samples is presented in Figure 7.7. In Figure 7.8, the segmented face images (using MTCNN, see Section 7.5.2) of each session and device are presented.

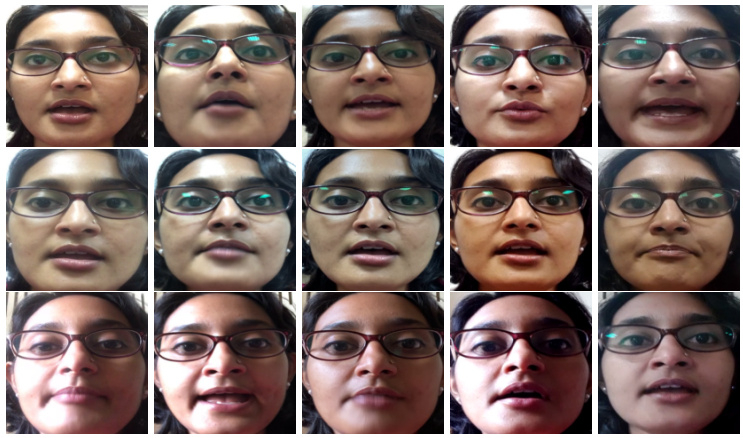


Figure 7.8: Detected face using MTCNN for face recognition. Left to Right: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8. Top row: Session 1, middle: Session2, bottom: Session3.

7.4.4 Presentation Attacks

We have created two types of presentation attacks: replay attacks and synthesized attacks.

Replay Attacks

The replay attacks are created by synchronized capture of audio-visual playback using Dell office monitor and Logitech speakers recorded on Samsung S8 phone.

Figure 7.9 show the replay attacks samples created in this work. The spectrograms of audio replay attacks are presented in figure 7.10.

Figure 7.9: Replay attack data sample. Left: Bona fide, right: Replay attack.

Synthesized attacks

Deep learning has been successfully applied to solve complex problems ranging from big data analysis to computer vision tasks and human level control. Advanced deep learning concepts have also been used to create threats to privacy, democracy and national security. One such deep-learning based application that loomed recently is "deepfake" (derived from 'deep learning' and 'fake'). For creating synthesized attacks, we have used deepfake approaches in this work. One of the approaches for creating face deepfakes is a technique where the face image of a source person is superimposed onto a target person to create a video/image of the target person. In this direction, the face-swapping model is proposed by Nirkin *et al.* [178] where swapping of face images are done in three stages. Reenactment and face segmentation is carried out in the first stage, followed by in-painting and blending. Reenactment, face transfer, or puppeteering uses facial expressions and

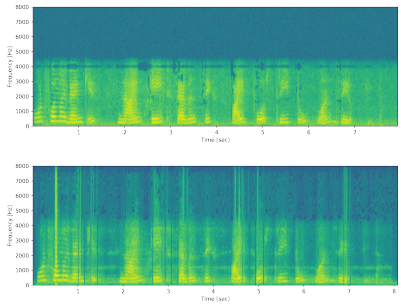


Figure 7.10: Spectrograms of bona fide and corresponding replay attack audio. Top: Bona fide, bottom: Replay attack.

assists in transforming the face in one video to guide the motions and deformations of the face appearing in another video or image. Face segmentation is performed using U-Net [179] and reenactment is performed using generative model named pix2pixHD [180]. In the second step, the occluded regions of the source face are mitigated using the same in-painting generator [180]. In the last step, a Gaussian Poisson Generative Adversarial Network (GP-GAN) [181] is used for high-resolution image blending for combining the gradient and colour information.

In our work, we have utilized FSGAN for swapping similar faces [8]. The face-swapping approach preserves the context of the target video by digitally overlaying the source's face landmarks. Therefore, the target video contains the key biometric characteristics of the source subject, which can efficiently be used as a presentation attack for the source's identity. Multiple deepfake datasets in the literature [182, 183, 184, 185] used a manual selection of faces for swapping. However, we have employed an automatic way to find a pair of similar faces in this work. We used cosine similarity of ArcFace embeddings to find a similar face for each of the male and female subjects (more on ArcFace in section 7.5.2). We have generated 97 face swapped videos for sentence 6 of bona fide data from session1 data of the Samsung S8 device.

WaveNet vocoder is used to generate high-quality raw speech samples conditioned on acoustic features [186]. The WavNet-based vocoder is popularly used in ASVSpoof 2019 challenge to create logical access presentation attacks [40]. In our work, we have used MFCC features as acoustic features in synthesizing 16-bit raw audio. We have adapted the implementation of WaveNet vocoder form the github[9] and pre-trained models from LJSpeech [187]. The figures 7.11 and 7.12

[8]FSGAN: https://github.com/YuvalNirkin/fsgan
[9]WaveNet Vocoder: https://github.com/r9y9/wavenet_vocoder

show the images samples and spectrograms of synthesized attacks respectively.

Figure 7.11: Face swap using FSGAN. Left: Source face, middle: Target face, right: Swapped face.

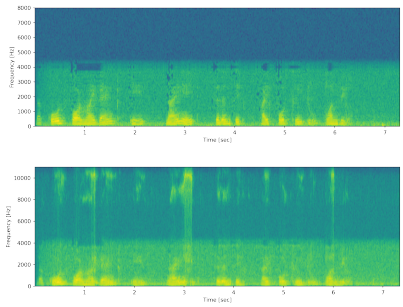


Figure 7.12: Spectrograms of bonafide and corresponding wavenet-vocoder synthesized audio. Top: Bona fide, bottom: Synthesized audio.

7.5 Performance Evaluation Protocols

The dataset is benchmarked with various face recognition, speaker verification and presentation attack detection methods. In this section, we explain briefly the baseline biometric systems employed along with evaluation metrics.

7.5.1 Automatic speaker Verification

I-vector based speaker Verification

The I-vector based ASV method is a Joint Factor Analysis (JFA) approach proposed in [188]. It models the channel effects and also speaker voice characteristics. The speech sample is represented as a low-dimensional super vector called i-vector. The i-vector represents the total factor in a speech utterance, including channel compensation which is carried out in a low-dimensional total variability space.

X-vector based speaker Verification

The deep neural networks (DNN) and end-to-end speaker verification approaches are state-of-the-art research methods that overcome handcrafted methods' drawbacks. The x-vector based speaker verification is a recent approach showing promising results in automatic speaker verification [32]. This method uses deep neural network (DNN) embeddings as features. The variable-length speech utterances are mapped to a fixed low-dimensional embedding (called x-vectors), and a deep network is trained to differentiate speakers. The training process requires a large amount of training data. Therefore, data augmentation is used along with added noise and reverberation to increase training data size. The implementations in Kaldi are employed in our work, and the pre-trained Universal Background Models, i-vector extractor and x-vector extractor are adapted to our experiments [10]. Probabilistic linear discriminant analysis (PLDA) [189] is used as a classifier for the i-vectors and x-vectors of enrollment and test samples. The log-likelihood score is computed between the enrolled and test speech sample pair.

Dilated residual network (DltResNet)

Extended ResNet implementation from [190] named dilated residual network (DltResNet) is used as the third speaker verification methods. The implementation is publicly available[11]. The DltResNet model is one of the state-of-the-art systems on the Voxceleb1 database evaluations achieving 4.8% EER on the dataset. The Euclidean distance between the DltResNet features is used for obtaining scores between enrolled and test samples.

[10] Kaldi GitHub: https://github.com/kaldi-asr/kaldi

[11] DltResNet: https://www.idiap.ch/software/bob/docs/bob/bob.learn.pytorch/v0.0.4/guide_audio_extractor.html

7.5.2 Face recognition

Face Detection

Face detection is performed as a prepossessing step on the video frames to detect and crop the face image. We have employed multitask cascaded convolutional networks (MTCNN) approach from Zhang *et al.* [191] for efficient face detection. The face recognition and face PAD methods used in this work used segmented face images.

Local Binary Patterns (LBP)

Local Binary Patterns (LBP) are a textual operator that labels the pixels in a face image according to neighbouring pixels' values and assigns a binary number. LBP for an image is calculated by assigning 0 or 1 to the pixel depending on the neighbour's pixel having high or low value. The resultant binary test is stored in an 8-bit array and later converted to decimal. This thresholding process, accumulating binary strings, and storing the decimal value is repeated for every pixel in the input image. Further, the LBP histogram is computed over the LBP output array. For a block, one of the $2^8 = 256$ possible patterns is possible. The advantage of LBP features is high discriminative power, computational simplicity, and invariance to grey-scale changes. LBPs have shown a prominent advantage in face recognition approaches. We used LBP histograms as features for face images and cosine distance to compute the score between the enrolled and test samples.

FaceNet face embeddings

The deep learning approaches have evolved into image processing and pattern recognition applications. In face recognition methods, FaceNet embeddings displayed an excellent image representation for facial features [24]. This is a deep face recognition approach that adapted the ideas from [192]. In this work, we have used the pretrained model on the VGGFace2 dataset using Inception ResNet v1. This model displayed an accuracy of 99.65% on the Labeled Faces in the Wild (LFW) dataset [193]. We have obtained FaceNet embeddings [12] for face detected images in our dataset and used cosine distance between the samples to obtain the verification scores.

ArcFace face descriptor

ArcFace face features are proposed in [25] for the large scale face recognition with enhanced discriminative power. ArcFace features emphasize the loss function in deep convolutional neural networks (DCNN) for clear geometric interpretation of

[12] FaceNet: https://github.com/davidsandberg/facenet

face images. The proposed descriptor is evaluated over ten face recognition benchmarks, and results show consistent performance improvement. We have employed the ArcFace implementation provided in Github [13]. The training data contains cleaned MS1M, VGG2 and CASIA-Web face datasets. ArcFace face descriptors are computed over detected face images, and similar to other face recognition methods, we have used cosine distance as a classifier.

In addition to the face recognition, we have used ArcFace face embeddings to obtain similarity scores between subjects in creating attacks in FSGAN face swapped videos (see section 7.4.4).

7.5.3 Presentation Attack Detection (PAD)

Voice PAD

The PAD methods used to evaluate the attacks created using speech are chosen from the baseline methods in the ASVSpoof 2019 challenge [40]. The two baseline methods are available in ASVSpoof 2019 evaluation protocols. Features used in these two methods are based on cepstral coefficients in the front-end and Gaussian Mixture Models (GMM) in the back-end. Linear Frequency Cepstral Coefficients (LFCC) and Constant Q Cepstral Coefficients (CQCC) are two features used to represent speech samples.

The LFCC features are similar to the Mel-frequency cepstral coefficients (MFCCs), with filters placed linearly in the exact sizes. The initial approach of LFCCs is used for the detection of synthetic speech in [194]. In this work, we used LFCC features are extracted with a frame length of 25ms and a 20-channel linear filter bank. An LFCC feature comprises 19 cepstral coefficients, a zeroth coefficient, static, delta, and delta-delta coefficients. The CQCC features are extracted with the toolkit provided in ASVSpoof 2019. The maximum frequency is set to fs/2, where fs is the sampling frequency, and the minimum frequency is fixed at $fs/2/2^9$ $15Hz$ (where 9 is the number of octaves) [195]. The number of bins per octave is set to 96, and re-sampling is applied with a period of 16. The dimension of features is 29 coefficients along with zeroth, static, delta, and delta-delta coefficients.

The front-end provides the cepstral coefficients, which are used to train 2-class GMMs in the back-end. The training process is carried out on the bonafide and attack speech samples with 512-component GMM models. An expectation-maximization (EM) algorithm is employed in training with random initialization. For testing, the scores of samples are calculated from the log-likelihood ratio with the help of trained bona fide and the attack speech models.

[13] ArcFace: https://github.com/deepinsight/insightface

Face PAD

The face recognition PAD methods are chosen from the baseline methods used in smartphone dataset evaluation in [12]. The two best-performing methods from five baseline methods are taken for evaluation in this work. These methods utilize local binary patterns (LBP) [196] and color texture features [197]. The support vector machines (SVM) are trained for different attacks and test for attack detection.

The LBP features are experiments for PAD in [196] for face attacks in a full biometrics verification system. In [38], the LBP features displayed a consistent performance of detecting attacks in different protocols of smartphone biometric data. Similarly, the experiments using colour texture features [197] resulted in the best-performing face PAD on smartphone face images. Therefore, we have included these methods in our evaluation of detection attacks.

7.5.4 Performance Metrics

The performance evaluation metrics from ISO/IEC [85] are utilized in our experiments to present and compare the results of different methods.

Verification Metrics

- False Match Rate (FMR) is the proportion of the completed biometric non-mated comparison trials that result in a false match.

- False Non-Match Rate (FNMR) is the proportion of the completed biometric mated comparison trials that result in a false non-match.

In addition to ISO/IEC metrics mentioned above, we have also presented an equal error rate (EER) to represent FMR and FNMR metrics in a single value. EER is the error rate at the point where FMR and FNMR are equal.

Presentation Attack Detection Metrics

- Impostor-Attack Presentation Match Rate (IAPMR) is the proportion of impostor attack samples (replay attacks) that are matched with bona fide samples. To compare ASV methods' performance, we have fixed FMR at 0.1% and presented FNMR and IAPMR for zero-effort impostors and attacks, respectively.

- Attack Presentation Classification Error Rate (APCER) is the proportion of attack presentations that are incorrectly classified as bona fide presentations, and Bonafide Presentation Classification Error Rate (BPCER) is the ratio of bona fide presentations incorrectly classified as attacks. This work presents

the BPCER_5 and BPCER_10 of PAD methods: the BPCER values at APCER are 5% and 10%, respectively.

Table 7.2: Inter-session speaker recognition evaluation (EER%).

Inter-session	i-Vector			X-Vector			DltResNet		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
S1	**5.31**	11.52	10.35	**5.31**	11.18	10.84	**4.85**	10.69	9.56
S2	11.70	**4.13**	10.51	11.20	**3.51**	9.96	10.63	**4.32**	9.50
S3	10.48	10.65	**5.16**	10.70	9.96	**5.23**	9.51	9.59	**4.53**

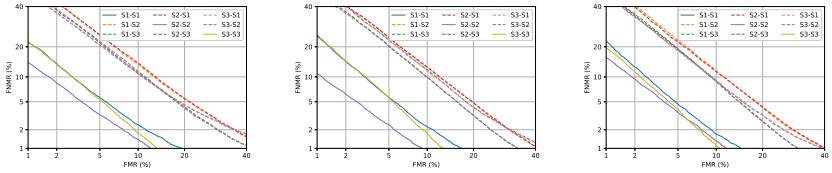


Figure 7.13: DET curves of inter-session speaker recognition experiments. Left: i-vector, middle: X-vector and right: DltResNet.

Also, we used Detection Equal Error Rate (D-EER) to present PAD methods' performance, a single value representation of APCER and BPCER. The score distributions of bona fide, zero-effort impostors and attacks are plotted along with the threshold of FMR = 0.1% to observe the impact of presentation attacks. Detection error trade-off (DET) curves plot the relationship between false match rate (FMR) and false non-match rate (FNMR) for bona fide samples or impostor attack presentation match rate (IAPMR) for attack samples, respectively.

7.6 Experimental Results

The main focus of this dataset is to provide scope for developing generalized biometric algorithms in face and speech-based recognition. The generalizability of a biometric algorithm can be achieved by considering multiple dependencies like session variance, device dependency and language. Therefore, in our work, we have performed experiments to demonstrate how these dependencies affect the state-of-the-art face and speaker recognition algorithms mentioned in 7.5. The benchmarking of the dataset is carried out by performing different experiments and presenting the results.

7.6.1 Automatic Speaker Verification

Automatic Speaker Verification methods display variable performance depending on the channel used to acquire and the noise present in the audio samples. In the

following experiments, we have evaluated the performance of the ASV methods in correspondence to the session, device and language.

Inter-session speaker recognition

The MAVS dataset contains data from three different sessions as explained in section 7.4. We have examined the session dependency by performing the inter-session speaker recognition. In this process, we have used the samples from one session to enrol and each of the other sessions to test. Table 7.2 presents the EER values displaying the comparison of three ASV methods on inter-session experiments.

- Session 2 data contains an added noise in all data samples. Therefore, it is seen that higher EER values are observed in all the results where session 2 data is used to enrol.
- However, when the same noise is present in test data, the ASV methods tend to perform better than the session with clean data (session 1). This concludes that ASV methods characterize the noise in the data and use it for recognition.
- Similarly, session 3 contains natural noise, which is not consistent in all samples, but it helps recognise the speaker better than the data with no noise.
- Alongside, DltResNet based ASV method displayed better performance compared to other methods.

Inter-device speaker recognition

Table 7.3: Inter-device speaker recognition evaluation (EER%) on i-vector method.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**1.86**	5.76	6.67	15.46	14.37
iPhone 10	5.88	**1.62**	4.74	15.02	13.97
iPhone 11	6.73	4.67	**1.47**	15.90	14.76
Samsung S7	15.51	14.90	15.70	**10.01**	13.26
Samsung S8	14.51	13.98	14.78	13.34	**8.77**

The properties of the data capturing device are key attributes for speaker recognition [188]. Although state-of-the-art ASV methods accommodate the channel characteristics, the change in devices from enrollment to test can still affect the speaker recognition performance. Our dataset used five different smartphones in

Table 7.4: Inter-device speaker recognition evaluation (EER%) on x-vector method.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**1.45**	5.82	6.55	15.33	14.09
iPhone 10	5.85	**1.81**	4.37	13.56	12.37
iPhone 11	6.54	4.30	**1.81**	14.27	13.10
Samsung S7	15.50	13.69	14.13	**8.55**	12.97
Samsung S8	14.04	12.25	12.93	13.30	**7.37**

Table 7.5: Inter-device speaker recognition evaluation (EER%) on DltResNet method.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**2.08**	6.52	7.07	16.56	16.38
iPhone 10	6.62	**2.03**	4.09	15.00	15.66
iPhone 11	7.06	4.03	**2.02**	15.92	16.14
Samsung S7	16.68	15.07	15.83	**7.04**	10.44
Samsung S8	16.51	15.52	16.11	10.63	**7.73**

data collection to examine the dependency of the device on ASV methods. Tables 7.3, 7.4, 7.5 show the EERs of all device combinations of enrollment and testing from the three ASV methods.

The results from inter-device experiments output some key points. These observations conclude the impact of channel dependency on state-of-the-art speaker recognition methods.

- The DltResNet method gave out the highest EER in most of the combinations even though it worked better with noisy data as shown in Section 7.6.1.
- The DNN based X-vector methods performed better than other methods.
- It is observed that the combinations of smartphones from the same manufacturer (Apple or Samsung) correlate with speaker recognition. When the enrollment and testing data are from the same manufacturer, the speaker recognition performs better than the cross-manufacturer combination.

Inter-language speaker recognition

The language difference in the audio sample for ASV has been a hot topic in recent years. Although there are datasets with utterances of the same person in different languages, the problem of language dependency is not benchmarked [12]. The degradation of biometric recognition due to language mismatch is presented in some

Table 7.6: Inter-language speaker recognition evaluation (EER%).

Inter-language	i-vector			x-vector			DltResNet		
	English	Hindi	Bengali	English	Hindi	Bengali	English	Hindi	Bengali
English	**5.47**	5.50	6.72	**4.98**	5.55	6.93	**4.88**	5.26	6.27
Hindi	5.58	**4.16**	5.33	5.45	**4.0**	5.60	5.32	**3.95**	5.15
Bengali	6.78	5.92	**5.08**	6.93	5.67	**5.21**	6.34	5.19	**4.87**

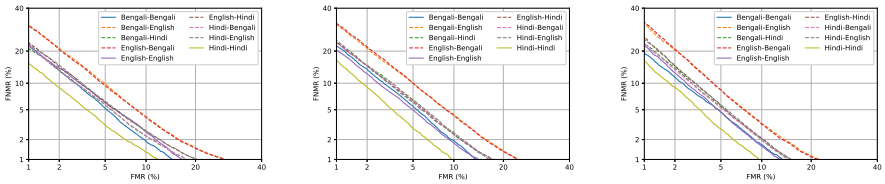


Figure 7.14: DET curves of inter-language speaker recognition experiments. Left: i-vector, middle: X-vector and right: DltResNet.

previous works [198], [199], [154]. Our dataset comprises of the same subjects speaking three different languages, therefore, providing scope for inter-language speaker recognition evaluation. Table 7.6 shows the inter-language speaker recognition evaluations.

- The problem of language mismatch from enrollment to testing is observed in all three ASV methods.
- However, the drop in EER is not high, but it is consistent across all the methods.
- It is important to notice that the training dataset contains multiple languages, and we assume that the extracted features contain language factors.
- Therefore, in the scenario of a small subset of languages in training data, the language mismatch problem would be considerable.

7.6.2 Face Recognition

The robustness of face recognition algorithms in smartphones is evaluated in this section. Similar to speaker recognition, we have performed two dependency experiments, namely inter-session and inter-device. The three face recognition systems are examined in these experiments by taking 20 equally distributed frames in each video.

Table 7.7: Inter session face recognition evaluation EER(%).

Inter-session	LBP			FaceNet			Arcface		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
S1	**5.39**	24.28	44.73	**0.26**	0.89	2.22	**3.42**	6.68	5.60
S2	24.28	**6.81**	41.55	0.87	**0.24**	1.65	6.42	**4.34**	6.81
S3	44.67	41.43	**4.43**	2.21	1.63	**0.21**	5.59	6.81	**1.43**

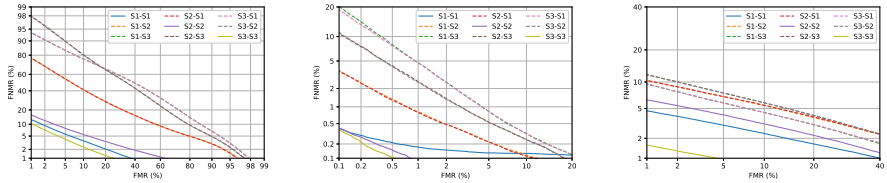


Figure 7.15: DET curves of inter-session face recognition experiments. Left: LBP, middle: FaceNet and right: ArcFace.

Inter-session

The session variability in face recognition is observed in this experiment.

- Session 2 and session 3 data has non-uniform lighting on the face region. Therefore, the cross-session face recognition displayed a clear drop in the performance.
- FaceNet performed better in attributing the problem of session variability among the three face recognition methods while displaying near-zero error rates in the same session.
- Table 7.7 present the EER values for inter-session face recognition experiments.

Table 7.8: LBP face recognition performance EER(%) in inter-device scenario.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**6.96**	19.50	19.60	22.94	31.21
iPhone 10	19.55	**5.32**	18.72	31.69	37.95
iPhone 11	19.70	18.76	**5.09**	25.67	32.60
Samsung S7	22.96	31.69	25.70	**5.05**	21.04
Samsung S8	31.13	37.87	32.65	21.10	**5.04**

Table 7.9: FaceNet face recognition performance EER(%) in inter-device scenario.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**0.20**	0.44	0.64	0.66	0.48
iPhone 10	0.45	**0.28**	0.51	0.69	0.53
iPhone 11	0.64	0.51	**0.3**	0.92	0.71
Samsung S7	0.67	0.68	0.90	**0.25**	0.34
Samsung S8	0.49	0.54	0.71	0.33	**0.16**

Table 7.10: Arcface face recognition performance EER(%) in inter-device scenario.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**3.30**	4.14	4.03	4.79	4.36
iPhone 10	4.10	**3.10**	3.76	4.76	4.31
iPhone 11	4.04	3.79	**3.01**	4.60	4.03
Samsung S7	4.80	4.76	4.55	**2.98**	3.78
Samsung S8	4.39	4.30	4.03	3.78	**2.72**

Inter-device

The results from inter-device experiments on face recognition are shown in Tables 7.8, 7.9, 7.10.

- The LBP features based face recognition displayed a high dependency on devices. When the device is the same in enrollment and testing, LBP features performed better face recognition. However, the recognition error has increased by three times when there is a miss-match in devices.
- Another observation is that the change in device manufacturer has also impacted face recognition similar to speaker recognition.
- FaceNet has displayed better face recognition considering the problem of device dependency. The drop in performance is observed, but it is not as consistent as other methods.
- ArcFace performed similarly to FaceNet in an inter-device face recognition scenario.
- Although the EER is higher in ArcFace than FaceNet; the device mismatch has not impacted the performance very much.

7.6.3 Audio-Visual Speaker Recognition

The audio-visual speaker recognition is performed by score-level fusion of best-performing face recognition and speaker recognition methods, FaceNet and X-vector methods, respectively. The score fusion approach used in this work is a simple averaging of scores obtained in individual verification methods.

Inter-session variance

Table 7.11: Inter session Audio-Visual speaker recognition evaluation EER(%).

Inter-session	S1	S2	S3
S1	**4.99**	10.73	10.46
S2	10.74	**3.21**	9.56
S3	10.34	9.55	**4.90**

- The combination of audio and visual data displayed similar results as that of individual biometric algorithms. This is because of the simple score-level fusion method employed in our work.
- We assume that an adaptive fusion approach would improve the performance.
- However, it introduces a new dependency on biometric algorithms in the form of a fusion approach.
- Table 7.11 show the results of inter-session audio-visual fusion experiments. Figure 7.16 present the corresponding DET curves.

Inter-device

The inter-device experiments on audio-visual biometric recognition are carried out similar to the inter-session approach. The obtained results display the same observations as that of audio-visual inter-session biometric recognition. It is clear from these experiments that an efficient fusion approach is required to take advantage of bi-modal biometrics. Table 7.12 display the EER values of inter-device experiments using audio-visual fusion.

7.6.4 Vulnerability from Presentation Attacks

The vulnerability of biometric recognition towards presentation attacks is examined in this section. The two types of presentation attacks created in this work are explained in Section 7.4.4. The biometric recognition performance before and after

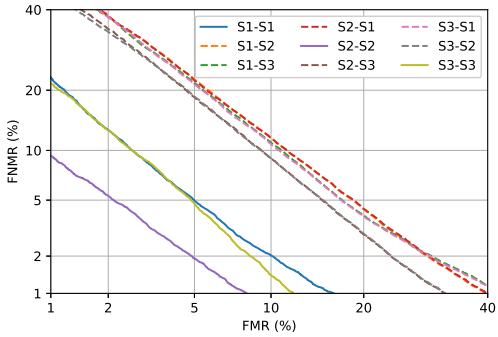


Figure 7.16: DET curves of inter-session experiments on Audio-Visual fusion of FaceNet and X-vector methods.

Table 7.12: Inter-device performance (EER%) of score-level fusion of FaceNet and X-vector methods.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**1.31**	5.53	6.24	14.55	13.30
iPhone 10	5.57	**1.65**	4.18	12.82	11.74
iPhone 11	6.25	4.15	**1.70**	13.53	12.41
Samsung S7	14.75	12.93	13.34	**7.92**	12.30
Samsung S8	13.28	11.54	12.30	12.59	**6.81**

the attacks is compared to check the robustness. When a presentation attack is not carried out, the performance is expressed in false non-match rate (FNMR) caused by zero-effort impostors. In presentation attacks, the vulnerability is presented as impostor attack presentation match rate (IAPMR).

Replay Attacks

The replay attacks are created by replaying an audio-visual biometric sample on a display and loudspeaker combination. The playback sample is recorded on one of the smartphones, namely the Samsung S8. The audio and face channels of replay attacks are examined for vulnerability individually on the two best performed biometric methods from the previous sections. For face recognition, FaceNet features are used, and for speaker recognition, X-vector features are employed.

- The impact of replay attack is presented in Table 7.13 in FNMR and IAPMR rates for zero-effort impostors and replay attacks, respectively.
- In face recognition, the vulnerability is observed as 96.87% IAPMR, repres-

Table 7.13: Replay attack vulnerability on Face and Voice at FMR = 0.1%

Biometric Algorithm	Zero-Effort impostors	Replay Attacks
	FNMR	IAPMR
FaceNet	0.09%	96.87%
X-vector	6.4%	25.93%

enting the number of attacks being matched with bonafide samples.

- The speaker recognition method displayed 25.93% IAPMR when compared to 6.4% FNMR.
- The score distributions of bona fide, zero-effort impostors and replay presentation attacks are presented in Figures 7.17 and 7.18.

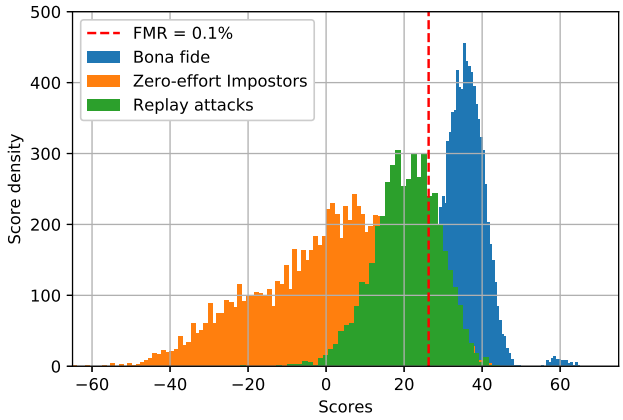


Figure 7.17: Audio Replay attacks score distribution tested on X-vector method.

Synthesized Attacks

Synthesized attacks are logical access attacks where the attack sample is presented digitally to the biometric system. Table 7.14 shows the vulnerability of synthesized attacks on face and voice modalities.

- The vulnerability evaluation on FaceNet based face recognition shows a 38.77% IAPMR, and the score distributions are presented in Figure 7.19.

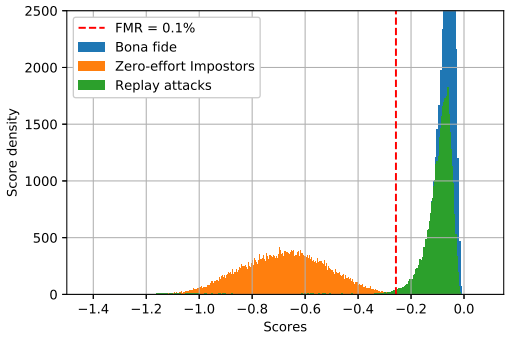


Figure 7.18: Video Replay attacks score distribution tested on FaceNet method.

- The speech synthesis is carried out using wavenet-vocoder, and the attacks displayed 99.68% IAPMR.
- The score distributions are presented in Figure 7.20.

Table 7.14: Synthesized attack vulnerability on Face and Voice at FMR = 0.1%

Biometric Algorithm	**Zero-Effort impostors**	**Synthesized Attacks**
	FNMR	IAPMR
FaceNet	0.21%	38.77%
X-vector	5.59%	99.68%

Audio-Visual Presentation Attacks

The vulnerability of audio-visual presentation attacks is examined with the help of fusion of presentation attacks on AV recognition methods explained in Section 7.6.3. The replay attacks and synthesized attacks are performed in individual biometric modalities, and the attack scores are fused to calculate the final scores. The impact of the audio-visual attacks is presented in Table 7.15 on two different attacks. Unlike unimodal biometric matching, the results of audio-visual biometrics are presented in False Rejection Rate (FRR) because it represents the system-level performance. Similarly, the score distributions are shown in Figures 7.21, 7.22.

- The results indicate that audio-visual fusion is vulnerable to presentation attacks.
- The problem of replay attacks is less compared to the synthesized attacks.

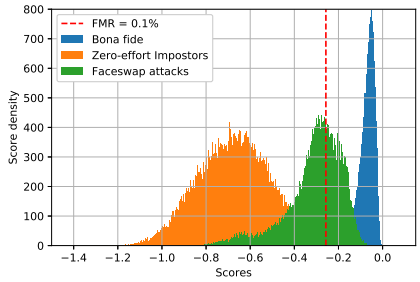


Figure 7.19: Score distribution of face swap attacks.

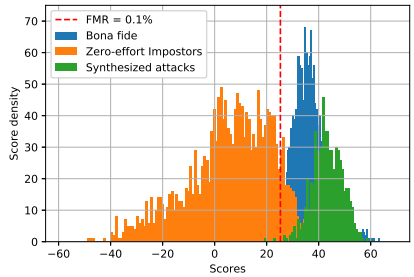


Figure 7.20: Score distributions of wavenet speech synthesized attacks.

- Although the replay attacks on face recognition displayed the highest vulnerability; the AV fusion approach appears to have the ability to overcome this problem. However, a similar observation is not seen in synthesized attacks.

- Thus, the AV fusion recognition approach has the vulnerability due to combined AV presentation attacks.

Table 7.15: Audio-Visual replay attacks vulnerability on AV fusion method at FMR = 0.1%

Attack Type	**Zero-Effort impostors**	**Presentation Attacks**
	FNMR	IAPMR
Replay Attacks	5.29%	28.46%
Synthesized Attacks	4.64%	99.83%

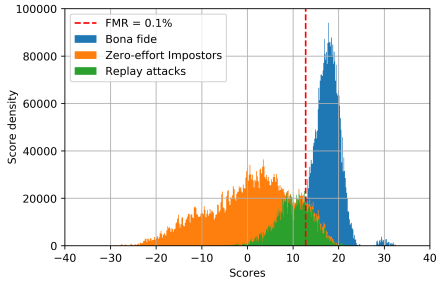


Figure 7.21: Audio-Visual replay attacks score distribution.

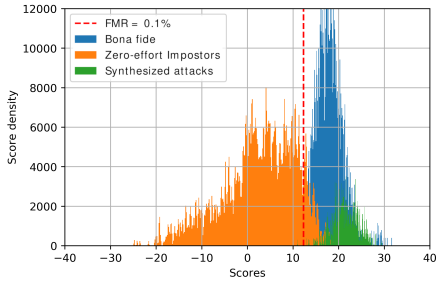


Figure 7.22: Audio-Visual synthesized attacks score distribution.

7.6.5 Presentation Attack Detection

The presentation attack detection experiments are performed using baseline PAD methods. The attack data is partitioned into three sets: training, developing and testing, with 35%, 35% and 30% of bona fide and attack samples, respectively. Each partition includes data from a unique set of subjects. We have chosen the baseline approaches used in Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVSpoof) for speaker recognition PAD in 2019. See Section 7.5.3. For face recognition, we opted the two best-performing methods from the face PAD methods used in [12]. Tables 7.16 and 7.17 show the results

Table 7.16: Results of speaker recognition presentation attack detection.

Attack type	LFCC-GMM			CQCC-GMM		
	D-EER	BPCER_5	BPCER_10	D-EER	BPCER_5	BPCER_10
Replay Attacks	44.14%	100%	93.15%	20.49%	45.63%	36.89%
Speech Synthesis	14.00%	39.82%	20.38%	14.08%	40.77%	22.33%

Table 7.17: Results of face recognition presentation attack detection.

Attack type	LBP-SVM			Color texture-SVM		
	D-EER	BPCER_5	BPCER_10	D-EER	BPCER_5	BPCER_10
Replay Attacks	4.96%	5.07%	1.28%	2.15%	1.35%	0.32%
FaceSwap	2.99%	1.74%	1.15%	2.54%	0.83%	0.26%

of the PAD methods in terms of D-EER, BPCER at APCER = 5% and BPCER at APCER = 10%. The DET curves in figures 7.23 and 7.24 present the performance of PAD methods.

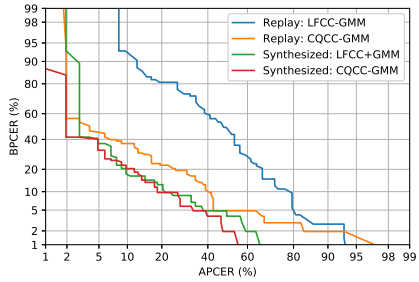


Figure 7.23: DET curves of voice PAD evaluation using baseline methods.

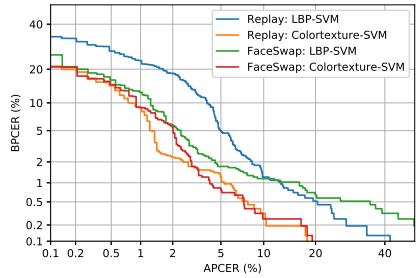


Figure 7.24: DET curves of face PAD evaluation using baseline methods.

- The voice PAD results indicate that the baseline methods are not able to detect the attacks.
- Alongside, replay attacks are difficult to detect when compared to synthesized attacks. In contrast, both face PAD methods performed well in detecting the attacks.

- The voice PAD methods are tested on the whole speech sample, where the face PAD methods are performed on detected face images in individual frames.
- Therefore, it is reasonable to assume that this could be the reason for the difference in performance.

Multimodal PAD

The presentation attacks on both modalities are possible with sophisticated equipment. The PAD methods should be able to detect the attacks before the verification process. In this experiment, we have fused the PAD scores from the CQCC-GMM method and the Color texture-SVM method to compute multimodal PAD scores. We have used a sum rule based fusion to combine two PAD methods. The table 7.18 shows the results of multimodal PAD approach and Figure 7.25 shows the PAD performance on two different types of attacks.

Table 7.18: Results of audio-visual PAD methods.

Attack type	Fusion PAD		
	D-EER	BPCER_5	BPCER_10
Replay Attacks	16.99%	38.83%	30.10%
Synthesized	11.87%	32.04%	15.54%

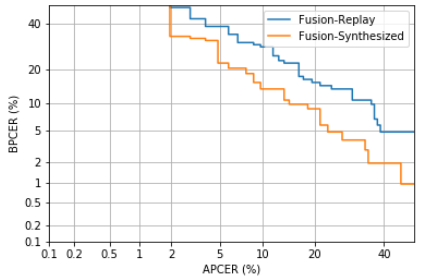


Figure 7.25: DET curves of audio-visual PAD of CQCC and Color texture methods.

- The replay attacks are observed to be difficult to detect compared to synthesized attacks. The performance of multimodal PAD is similar to individual PAD in regards to the types of attacks.
- The multimodal PAD does not improve the attack detection performance. The reason for this could be the usage of simple sum rule based fusion.

- The co-related and complementary information between audio and visual domains is not taken into account in this fusion approach. Therefore, multimodal PAD does not show any promising improvement over individual PAD approaches.

7.7 Conclusion

Smartphone biometrics have emerged into advanced security applications like banking transactions and identity verification. The built-in biometric systems by smartphone manufacturers can be utilized for this purpose. However, it is difficult to entirely rely on the built-in systems due to the variance in sensors and unknown algorithms embedded into smartphones. In this direction, it is possible to use the default sensors in smartphones like cameras and microphones. Therefore, we have developed a multidimensional smartphone audio-visual dataset that includes different languages, devices, sessions, and texts in this work. We have presented in this paper some of the previous works on building an audio-visual dataset and discussed our multi-lingual smartphone audio-visual (MAVS) dataset.

Further, we have performed experiments on examining the robustness of state-of-the-art biometric algorithms in two directions. The first direction concerns the problem of algorithm dependencies that include signal noise, capturing device and speech language. We have prepared inter-session, inter-device and inter-language experiments and presented the results. In the second direction, presentation attacks are evaluated for the vulnerability of biometric algorithms and the performance of baseline PAD algorithms. The results show the requirement of robust audio-visual biometrics algorithms to deal with the problems of multiple dependencies and presentation attacks. The proposed dataset would help the research community in developing advanced biometric algorithms and presentation attack detection approaches.

7.7.1 Future work

The MAVS dataset is made publicly available for research purposes [14]. The proposed dataset can be used in multiple directions in smartphone audio-visual research. The future work in this research direction using the dataset is as follows.

1. Novel biometric algorithms are modelled by identifying various problems that question the robustness of smartphone authentication.

[14] MAVS dataset request form: https://docs.google.com/forms/d/e/1FAIpQLSfTMqnQj8KNoUi1Ms1tx8Ewgil2l4wAAJVaKUJs6VkWfjAo4w/viewform?usp=sf_link

2. The authentication technology through biometrics can be improved via Audio-visual person recognition through the efficient usage of complementary information between audio and visual modalities.

3. The dataset contains subjects of different ages ranging from 18 to 48 years and gender labels (70 male and 33 female). Therefore, the dataset can be used for studying gender classification and fairness. Further, the audio data from three different languages can be used for language detection.

4. The correlated information between biometric cues are used to propose advanced presentation attack detection algorithms towards unknown and unseen attacks. E.g. lip-sync, correlated biometric data.

5. Generalizable biometric algorithms are developed in smartphone environments for real-world applications across different devices and capturing conditions.

Acknowledgment

We acknowledge the Idiap Research Institute and Prof. Sébastien Marcel for the data capture mobile application developed as a part of the SWAN (Secured access over Wide Area Network) project funded by the Research Council of Norway (Grant No. IKTPLUSS 248030/O70).

Chapter 8

Article 3: Image Quality and Texture-Based Features for Reliable Textured Contact Lens Detection

Mandalapu, Hareesh, Raghavendra Ramachandra, and Christoph Busch. Image quality and texture-based features for reliable textured contact lens detection. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 587–594.IEEE, 2018

8.1 Abstract

Textured contact-lens detection in iris biometrics has been a significant problem. In this paper, we propose a novel approach based on image quality and texture-based features for presentation attack detection for patterned/textured contact lens detection. The proposed approach employs the image quality features computed using Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) and texture features computed from Binarized Statistical Image Features (BSIF) to detect presentation attacks based on contact lenses. An efficient comparator using Spectral Regression Kernel Discriminant Analysis (SRKDA) is used for computing sample scores. The Fischer Discriminant Ratio (FDR) weighted fusion is used to perform score-level fusion from both models. The proposed method is tested on LivDet-Iris 2017 Clarkson, Notre Dame, and IIITD dataset. The experiments show noticeable results in detecting textured contact lens in iris samples for both same-set evaluation and cross-set evaluation.

8.2 Introduction

Iris is one of the unique and stable biometric characteristics and has been well used in several applications like secure border control and token free access control [26, 200]. The main advantage of the iris biometric characteristic is it's uniqueness and permanence when compared to other characteristics such face or fingerprint. The uniqueness can be observed even for mono-zygotic twins making it a preferred modality for robust biometrics. Further, unlike the face or fingerprint, iris is a protected biometric characteristic covered by eye-lashes making it not prone to environmental damages (for eg., cuts or abrasions in the fingerprint).

However, iris recognition systems are vulnerable due to several kinds of presentation attacks. Well-known presentation attacks against iris capture devices are printed iris image attacks, electronic display attacks (images/videos presented on screen), or textured-contact-lens attacks either to conceal iris identity or attack pre-enrolled iris images. Among the wide range of attacks, the patterned/textured contact lens attacks pose a severe threat to iris recognition [3] as it is difficult to be noticed by a supervising operator (e.g. a border guard) in case of assisted biometric access control. Due to the reason that contact lens overlays the iris itself and move along with eye ball, the patterns on the lens are difficult to distinguish from the actual iris pattern.

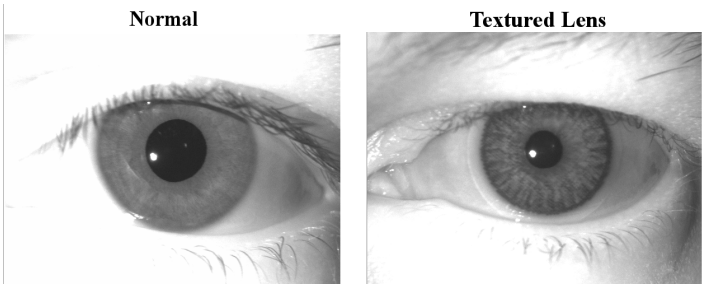


Figure 8.1: Example iris images with no lens and textured contact lens.

8.3 Related Work

Vulnerabilities in iris biometrics have been rapidly growing due to the availability of the technology to generate artefacts (i.e. presentation attack instruments) at low-cost. Among the different presentation attack instrument species, the use of textured/patterned contact lens can be considered as an efficient way to conceal the identity of the subject when interacting with an iris recognition systems [3]. The availability of low-cost patterned contact lens further makes the attacks to be carried out easily. It is therefore critical to detect the textured or patterned contact

lenses to prevent the attacks on iris systems.

Table 8.1: Texture analysis based contact lens attack detection algorithms.

Author	Feature extraction algorithm	Dataset
Gragnaniello *et al.* [201]	Scale-invariant image descriptor	Notre Dame and IIITD
Yadav *et al.* [3]	Local Binary Patterns (LBP) and Pyramid Histogram of Oriented Gradients (PHOG)	IIITD and NDCLD
He *et al.* [4]	Gray level co-occurrence matrices (GLCM)	SJTU iris database
Komulainen *et al.* [202]	Binarized Statistical Image Features (BSIF)	NDCLD'13
Raghavendra *et al.* [203]	Statistically independent filters for Binarized Statistical Image Features (BSIF)	IIITD and NDCLD'12
Doyle and Bowyer [204]	Binarized Statistical Image Features (BSIF)	NDCLD'15
Kohli *et al.* [205]	Multi-order dense Zernike moments and Local Binary Patterns (LBP) with variance	IIITD
Hu *et al.* [206]	Regional features via spatial pyramid and relational measure	Clarkson, Warsaw, Notre Dame and MobBIOfake

In the literature, a number of methods are proposed for classifying iris images into bona fide (i.e. real) and contact lens images (i.e. presentation attack samples). Xiaofu He *et al.* [4] used an statistical texture analysis based approach to detect artefact iris samples. Wei *et al.* [207] proposed three methods to detect attack iris images: using iris edge sharpness, iris texton-features and selected features from the co-occurrence matrix. He *et al.* used local binary patterns (LBP) on sub-divided images of the normalized iris image for iris presentation attack detection [208]. Zhang *et al.* used scale-invariant feature transforms (SIFT) descriptors and ranked local binary patterns (LBP) sequence, to compute weighted LBP map for detecting contact lens [209]. A novel texture pattern called Hierarchical Visual Codebook (HVC) was proposed by Sun *et al.* for sparse representation of the iris texture to classify artefact and bona fide iris images [210]. Scale-invariant descriptors on segmented iris images were used by Gragnaniello *et al.* to detect contact lenses [201]. Recent works in this direction used Binarized Statistical Image features (BSIF). In Komulainen *et al.*, applied BSIF features from pre-processed iris images were used to obtain more generalized results in detecting textured contact lens [202]. Statistically independent filters were used by Raghavendra *et al.* [203] to extract BSIF features. Doyle and Bowyer [204] used BSIF filters at multiple scales for detecting textured contact lenses. A brief overview of the state-of-the-art texture feature based contact lens presentation attack detection methods is presented in Table 8.1.

Despite the extensive state-of-the-art approaches targeted to detect patterned/textured contact lens, the generalization across different species of patterned/textured contact lenses and also across different capture devices is limited. In this work, we propose a novel method based on quality based features extracted using Blind/Referenceless

Image Spatial Quality Evaluator (BRISQUE) features [41] and texture-based features extracted using Binarized Statistical Image features (BSIF)[211]. In the next step, we use the Spectral Regression Kernel Discriminant Analysis (SRKDA) [43] independently on the BSIF and BRISQUE features to obtain the Presentation Attack Detection (PAD) scores. Finally, we employ the Fischer Discriminant Ratio (FDR) method to combine the detector scores to spot the patterned/textured contact lens attack reliably. The following are the main contributions of this paper:

- Novel approach for patterned/textured contact lens detection algorithm using image quality and texture-based features.
- Extensive experiments on three different publicly available datasets and comparison with state-of-the algorithms which have indicated the superiority of the proposed method.
- When compared to state-of-the-art techniques, the proposed method uses only the full image, corresponding to the ocular region, and thus the method overcomes the need of iris region segmentation or the strip image used in [203].
- New experiment protocol by dividing the whole dataset into three non-overlapping groups corresponding to development, training and testing.

The rest of the paper is organized as follows: Section 8.3 presents the existing texture-based presentation attack detection algorithms. Section 8.4 explains the proposed method based on image quality and texture-based features. The dataset used, experiments performed and results obtained during this work are presented in Section 8.5. In Section 8.6, the conclusion of this paper is presented.

8.4 Proposed Method

The primary objective of the proposed method is to reliably detect patterned contact lenses from iris images that are captured from different iris sensors. To this extent, we propose a novel approach that explores both image quality and texture features. The use of image quality features will play a vital role in achieving the generalization across the iris capture device and the use of texture features contribute in observing the texture information from the patterned contact lens manufactured from different vendors. Therefore, the proposed method is expected to achieve generalization capability across both captured devices and the patterned contact lens vendors.

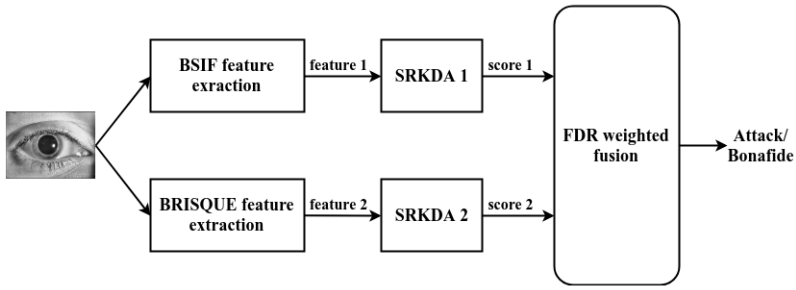


Figure 8.2: Block diagram of the proposed method.

Figure 8.2 shows the block diagram of the proposed approach that can be structured into three main functional blocks: (1) feature extraction block (2) comparison block (3) score level fusion block.

8.4.1 Feature Extraction

Given the input image I, the proposed method will extract both quality and texture feature using BRISQUE and BSIF features respectively.

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) Features

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is a natural scene statistic-based distortion-generic image quality assessment (IQA) model that operates in the spatial domain to provide the holistic measure of the image quality [41]. In this work, the BRISQUE features are extracted starting by measuring locally normalized luminance measures that are computed with the help of local mean subtraction and divisive normalization. Then, the spatial features are extracted by calculating generalized Gaussian distribution characteristics. The shape and variance coefficients of the image are obtained by fitting Mean Subtracted Contrast Normalization (MSCN) image to the Generalized Gaussian Distribution (GGD). The correlation images are obtained from four directions - horizontal, vertical and diagonal, in which the images are fitted using Asymmetric Generalized Gaussian Distribution (AGGD). Thus, given the image I, the extracted BRISQUE feature I_{BQ} will be of dimension 18×1. In this work, we used a feature vector of dimension 36×1 representing an image by concatenating two BRISQUE features, one from the original image and one from the image that is obtained by downsizing the original image to half of its original size.

Binarized Statistical Image Features (BSIF)

The texture-based features used in this work are computed using Binarized Statistical Image Features (BSIF) [202]. The use of BSIF features is known to be more

suitable for detecting patterned/textured contact lens [42]. The objective of BSIF is to use a filter based on ICA unsupervised scheme and represent each pixel in a binary code. The binary code is used as a local descriptor of the image. The histogram of the pixel values is used to characterize the texture properties of the image.

Given an image I and a filter F_i, the filter response is obtained as:

$$r_i = \sum_{m,n} I.F_i \tag{8.1}$$

where F_i , $\forall i = 1, 2, ..., n$ denotes the number of statistically independent filters whose response can be computed together and binarized to obtain the binary string as:

$$b = \begin{cases} 1, \text{ if } r_i > 0 \\ 0, \text{ otherwise} \end{cases} \tag{8.2}$$

The obtained BSIF features are used to calculate the histogram of pixel's binary codes that characterize the texture components in the image. In this work, we have used the filters of size 7×7 with a length of 10 bits which extract a rich set of information for presentation attack detection tasks. The choice has been made by considering the performance [212] and accuracy of iris presentation attack detection from the previous works [42].

8.4.2 Comparator: SRKDA

Spectral Regression Kernel Discriminant Analysis (SRKDA) was proposed in [43] to overcome the computational problems due to Eigen decomposition. SRKDA uses spectral graph analysis and casts discriminant analysis into a regression framework. To solve the optimization problem, Eigen decomposition is replaced by regression which results in significant improvement in speed compared to ordinary kernel discriminant analysis (KDA) methods. SRKDA provides efficient computation and regularization techniques which fully utilize the computational results of training samples. It has been proven that SRKDA has better performance than Linear Discriminant Analysis (LDA), Kernel Discriminant Analysis (KDA) and Support Vector Machines (SVM) [43]. In this work, we train the SRKDA classifier independently on the BRISQUE and BSIF features from the training dataset [213]. We have used the Gaussian RBF as a kernel function for SRKDA. Given the testing dataset, we obtain the comparison score corresponding to BRISQUE and BSIF and let these scores be represented as C_{BQ} and C_{BS} respectively.

8.4.3 Score-Level Fusion

In this work, we employ the Fisher Discriminant Ratio (FDR) [44] based weighted fusion algorithm to combine the comparison score corresponding to two different features. The FDR of an algorithm can be calculated as:

$$FDR_k = \frac{(\mu_k^G)) - (\mu_k^I)}{(\sigma_k^G)^2 + (\sigma_k^I)^2} \tag{8.3}$$

where μ_k^G and μ_k^I represent the mean values of genuine and impostor scores of algorithm k respectively and σ_k^G and σ_k^I are the standard deviations. The weight w_k of an algorithm k can be obtained as:

$$w_k = \frac{FDR_k}{\Sigma_{k=1}^{N} FDR_k} \tag{8.4}$$

The weighted score (S) of a sample on two algorithms, namely BRISQUE and BSIF, can be calculated as:

$$S = w_1 \times C_{BQ} + w_2 \times C_{BS} \tag{8.5}$$

Where, w_1 and w_2 denotes the weights that are computed using the Fisher Discriminant Ratio [44]. For more information on the Fisher Discriminant Ratio, readers can refer to [44]. Note that, the weights are computed only on the development set and applied on the testing set to evaluate the final performance.

8.5 Experiments and Results

In this work, extensive experiments were performed on the publicly available LivDet-Iris 2017 dataset [45]. We carried out two different types of experiments, to quantitatively evaluate and compare the performance of the proposed method and the state-of-the-art methods. **Experiment 1** evaluates the performance of the proposed method with the same dataset used for development, training and testing. **Experiment 2** evaluates the performance of the proposed method with cross-dataset scenario, such as one dataset is used for developing the PAD sub-system and another dataset is used for training and testing.

8.5.1 LivDet-Iris 2017 datasets

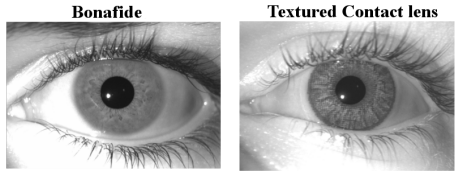
The LivDet-Iris 2017 dataset includes Clarkson, Notredam and IIITD dataset. In the following subsections, a brief description of each dataset is provided.

Clarkson Contact Lens dataset

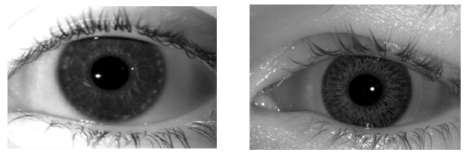
The Clarkson contact lens dataset is part of the Clarkson dataset, which contains live, print and patterned contact lens images. The Clarkson dataset was acquired using LG IrisAccess EOU2200 camera. The training set of Clarkson dataset consists of 2469 live images and 1122 patterned contact lens images. The live iris images were collected from 25 subjects and patterned contact lenses were collected from 5 subjects wearing 15 contact lens. The testing set consists of 1485 live images from 25 subjects and 765 patterned contact lens images from 7 subjects.

Notre Dame Contact Lens dataset

The Notre Dame contact lens dataset was captured using two sensors LG 4000 and AD 100. Unlike the LivDet-Iris competition, we kept the samples from different sensors separately. Both the training and testing set of LG 4000 sensor data contains 1000 live images and 400 textured contact lens each. The AD 100 sensor data contains 200 live images and 100 textured contact lens images in each of the training and testing tests. The textured contact lens used in this dataset were from five different manufacturers: J&J, Ciba, Cooper, UCL and ClearLab. Figure 8.3 shows the example images of live and textured contact lens from two sensors in Notre Dame dataset.



(a) LG 4000 sensor



(b) AD 100 sensor

Figure 8.3: Example images of bona fide samples and textured contact lens attack samples stemming from two sensors of the Notre Dame dataset

IIITD Contact Lens Iris (CLI) dataset

The IIITD contact lens dataset also contains textured contact lenses from various manufacturers: CIBA Vision Freshlook Dailies, Bausch and Lomb Lacelle, and Aryan 3-Tone. Two sensors: Cogent and Vista were used in capturing the images. The IIITD Cogent dataset contains 1143 live and 1138 textured contact lens im-

Table 8.2: Amount of training, development (Dev) and testing data samples used in this work. (B: bona fide, CL: contact lens)

Dataset	Train set		Dev set		Test set	
	B	CL	B	CL	B	CL
Clarkson	1646	748	823	374	1485	765
Notre Dame AD 100	133	133	67	67	100	100
Notre Dame LG 4000	666	666	334	344	400	400
IIITD Cogent	466	466	234	234	443	438
IIITD Vista	333	333	167	167	490	490

ages. 700 bona fide images and 700 textured contact lens images of the Cogent dataset were taken for training, development and rest (443 bona fide and 438 textured contact lens) are used for testing. The IIITD vista dataset contains 990 bona fide and 990 textured contact lens images. The vista dataset is divided into 500 bona fide images and 500 textured contact lens images for training, development and rest of the samples (490 bona fide and 490 textured contact lens) are taken for testing.

8.5.2 Performance Evaluation Protocol

To effectively evaluate the performance of the proposed method and the state-of-the-art methods, we divided each dataset to have three non-overlapping sets namely: development, training, and testing. The development dataset is used to tune the parameters of the proposed method, training dataset is used to train the PAD model, and also to set the operating threshold of the PAD systems at APCER = 5% and 10%. The testing set is solely used to report the performance results of both proposed PAD algorithm and the state-of-the-art PAD algorithms.

In this work, we present the results by following the International Standard ISO/IEC 30107-3 [37] in terms of Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER). APCER and BPCER metrics are described as follows:

- Attack Presentation Classification Error Rate (APCER) is the proportion of attack presentations that are incorrectly classified as bona fide presentations.
- Bona fide Presentation Classification Error Rate (BPCER) is the proportion

Table 8.3: Experiment 1: Results comparison of proposed method with state-of-the-art PAD methods [3] [4]. Training and testing on same dataset. D-EER(%), BPCER_5 is BPCER(%) at APCER=5% and BPCER_10 is BPCER(%) at APCER=10%.

Dataset	Error rate	LBP+SVM [3]	GLCM+SVM [4]	LBP+PHOG+SVM [3]	Proposed method
Clarkson	D-EER	42.75	33.10	32.40	**1.55**
	BPCER_5	81.54	68.95	73.8	**0.94**
	BPCER_10	74.34	58.51	61.21	**0.6**
Notre Dame AD100	D-EER	8	10	2	**0**
	BPCER_5	10	11	**0**	**0**
	BPCER_10	8	10	**0**	**0**
Notre Dame LG 4000	D-EER	27.25	4.25	2.75	**0**
	BPCER_5	45	3.75	2	**0**
	BPCER_10	40.25	3.5	1	**0**
IIITD Cogent	D-EER	39.72	41.99	18.72	**11.69**
	BPCER_5	92.09	88.26	54.62	**47.14**
	BPCER_10	79	73.81	33.86	**19.4**
IIITD Vista	D-EER	26.73	40.61	10	**6.32**
	BPCER_5	80.8	89.59	23.26	**10**
	BPCER_10	62.65	77.55	10.2	**3.85**

of bona fide presentations that are incorrectly classified as presentation attacks.

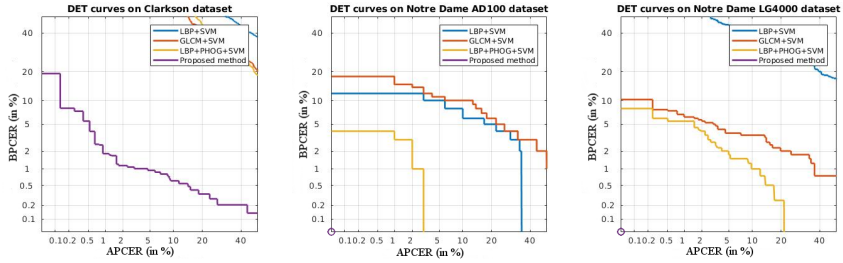
We computed two BPCER values for each dataset, BPCER_5 and BPCER_10, by fixing APCER at 5% and 10% respectively. In addition, we also include the results in terms of Detection Equal Error Rate (D-EER%).

8.5.3 Results and Discussion

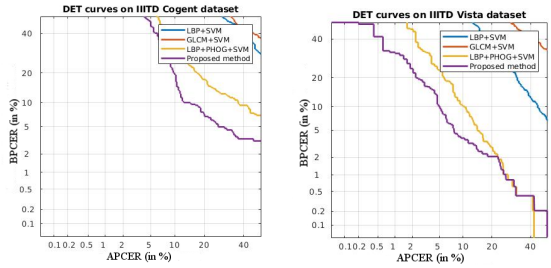
Experiment 1

Table 8.3 indicate the quantitative results of the proposed method in Experiment 1. The proposed method is compared to three state-of-the-art methods: (1) Local Binary Patterns(LBP) + Support Vector Machine(SVM) [3], (2) Gray-level Co-occurrence Matrix(GLCM) + Support Vector Machine(SVM) [4], (3) Local Binary Patterns(LBP) + Pyramid Histogram of Oriented Gradients(PHOG) + Support Vector Machine(SVM) [3]. Based on the obtained results, following are the main observations:

- The proposed method indicates an improved performance in accuracy across all the individual dataset. More particularly, the proposed method has indicated the best results on the Notre Dame AD100 and Notre Dame LG4000 datasets with D-EER (%) of 0%.



(a) Clarkson (b) ND AD100 (Proposed method D-EER = 0%) (c) ND LG4000 (Proposed method D-EER = 0%)



(d) IIITD Cogent (e) IIITD Vista

Figure 8.4: Detection Error Trade-off (DET) curves: Proposed method and the state-of-the-art methods [3] [4] from Experiment 1

- The proposed method has indicated the superior performance, when compared to that of the state-of-the-art techniques across all datasets. This will indicate the efficacy of the proposed method across the different datasets for patterned/textured contact lens detection.
- When compared to the performance of the proposed method across five different dataset, the degraded performance of the state-of-the-art together with the proposed method is noted for the IIITD Cogent sensor. This fact can be attributed to the quality of the images acquired using the Cogent sensor. However, it is interesting to note that, the proposed method has indicated the best performance across all datasets that can be attributed to the use of both quality and texture-based features that can allow the generalization across different types of patterned/textured contact lenses.
- Figure 8.4 shows the Detection Error Trade-off (DET) curves indicating the performance of the proposed method together with the state-of-the-art methods across all five different dataset which indicate the superior performance of the proposed method.

Table 8.4: Experiment 2: Cross dataset validation by comparison of results from proposed method with state-of-the-art PAD methods [3] [4]. Training on one dataset and testing on all other datasets combined. D-EER(%), BPCER_5 is BPCER(%) at APCER=5% and BPCER_10 is BPCER(%) at APCER=10%.

Training Dataset	Error rate	LBP+SVM [3]	GLCM+SVM [4]	LBP+PHOG+SVM [3]	Proposed Method
Clarkson	D-EER	50	45.96	39.04	**37.88**
	BPCER_5	95.88	94.39	**90.50**	91.34
	BPCER_10	91.62	89.95	**78.85**	83.53
Notre Dame AD 100	D-EER	50	50	50	**46.54**
	BPCER_5	97.26	96.02	94.28	**94.07**
	BPCER_10	92.93	92.08	91.87	**89.78**
Notre Dame LG 4000	D-EER	**43.26**	50	50	48.50
	BPCER_5	95.55	96.74	**95.23**	96.54
	BPCER_10	**90.03**	93.20	91.89	90.98
IIITD Cogent	D-EER	50	**44.84**	50	50
	BPCER_5	97.73	90.82	**84.40**	98.46
	BPCER_10	94.18	82.82	**81.37**	97.13
IIITD Vista	D-EER	50	49.02	50	**36.63**
	BPCER_5	97.28	96.08	97.15	**83.03**
	BPCER_10	93.24	93.9	95.05	**70.46**

Experiment 2

In this experiment, we analyzed the performance of the proposed method on cross-dataset validation and the results are presented in Table 8.4. The training set is fixed to one subset of data and other dataset combined are set as testing set. In this way, the generalization of the proposed method can be evaluated. The proposed method is compared to the state-of-the-art methods [3, 4] similar to the *Experiment 1*. The important observations of this experiment are:

- The proposed method gives an improvement in results over Clarkson, Notre Dame AD 100 and IIITD vista dataset. This shows that the proposed method performs better than state-of-the-art methods in cross-dataset evaluation.
- For the two datasets - Notre Dame LG 4000 and IIITD cogent, the difference in D-EER% of the best method and the proposed method is just over 5%. This shows the robustness of the proposed method to cross-dataset patterned/textured contact lens detection.
- In case of IIITD vista dataset, the proposed method display D-EER of 36.63% whereas the best D-EER among state-of-the-art methods is 49.02%. This shows that the proposed method gives a significant improvement of more than 13% in cross dataset textured contact lens detection of IIITD vista dataset.

8.6 Conclusion

Iris recognition is reported to be vulnerable to the textured contact lens presentation attacks that can be used for concealing the identity. The patterned contact lens detection is a challenging problem because of the variability in lenses that can be attributed to different manufacturers and different capture devices. In this paper, we presented a novel approach using image quality and texture-based features for the presentation attack detection algorithm of contact lens detection. The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) features provide image characteristics independent of the sensor and capturing conditions. The Binarized statistical image features (BSIF) provide texture information of iris image. We used a robust fusion of these features at score level that can lead to the reliable patterned contact lens detection. The proposed method is based on using the ocular image captured using the iris sensor and thus overcomes the computation of the iris detection and segmentation. Experiments are carried out on the publicly available dataset with sophisticated evaluation protocol that partition the dataset into three non-overlapping sets namely: development, training and testing. The obtained results have indicated the outstanding performance of the proposed PAD method when it is trained and tested with the same dataset. Additional experiments on the cross-datasets further demonstrate the limited robustness of all methods for generalization tasks. In the future work of this direction, we compare the efficiency of image quality feature BRISQUE with the standardized iris sample quality metrics from ISO/IEC 29794-6 [214].

Acknowledgement

R. Raghavendra and Christoph Busch are supported by Research Council of Norway (Grant No. IKTPLUSS 248030/O70).

Chapter 9

Article 4: Empirical Evaluation of Texture-Based Print and Contact Lens Iris Presentation Attack Detection Methods

Mandalapu, Hareesh, Raghavendra Ramachandra, and Christoph Busch. "Empirical evaluation of texture-based print and contact lens iris presentation attack detection methods." In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*, pp. 7-14. 2019.

9.1 Abstract

Iris-based identification methods have been popularly used in real-world applications due to the unique characteristics of iris when compared to other biometric characteristics like face and fingerprint. As technological advances and low-cost artefacts are becoming more available, vulnerabilities to iris biometrics due to presentation attacks (PAs) are becoming a challenging problem. Presentation attack detection (PAD) algorithms have been employed in biometric capture devices and it has been an active research topic in the past years. In this study, a detailed survey and evaluation of state-of-the-art texture-based iris PAD methods are performed. Five different PAD methods are tested on four different datasets consisting of print and contact lens presentation attacks. Extensive experiments are performed on four different scenarios of presentation attack and results are presented. The properties of PAD algorithms like the quality of the database, the generalization abilities are mainly discussed in this work. It has been observed that

fusion-based PAD methods perform better than other methods.

9.2 Introduction

The iris biometric characteristic is considered to be unique even for identical twins and it is a protected biometric characteristic covered by eyelids which makes it not prone to external damages (e.g., cuts or abrasions that occur to fingerprints). This makes the observation and recognition of iris patterns more robust than other modalities like face and fingerprint. Iris recognition systems have been used in real-world applications in many fields [26, 200]. Though computational power has evolved rapidly, an iris recognition system cannot assure the security that it is supposed to provide. The vulnerabilities of an iris capture device are becoming a provision for attackers to have unwanted authorization. Presentation attacks are one of the main reasons for vulnerabilities in biometrics. Recent growth in computational abilities and availability of low-cost artefacts made it easy to develop attacks which form a potential threat to biometric systems.

Well-known presentation attacks against iris capture devices are printed iris image attacks, electronic display attacks (images/videos presented on screen), textured-contact-lens attacks and synthetic /plastic eyes. Increasing development of iris biometrics in daily usage is making presentation attack detection a popular topic in research. The state-of-the-art PAD methods are highly restricted to the type of attacks and biometric systems that were used. For example, change in single attribute alters the performance of PAD (e.g. type of attacks, iris sensor and capturing conditions). This leads to the problem of generalization of PAD in detecting various kinds of presentation attacks in different scenarios. This paper focuses on performing a survey and evaluation of popular texture-based PAD methods that are developed in detecting two different kinds of attacks (print and contact lens) in different real-world scenarios.

The rest of the report is organized as follows: Section 9.3 discusses the state-of-the-art texture-based PAD algorithms. Section 9.4 presents the evaluation methodology used in this work which includes 5 PAD methods, 4 datasets and performance protocol. Section 9.5 explains the experiments performed during the evaluation and presents the results with discussions. Finally, in Section 9.6, conclusion of this work is given.

9.3 Related Work

There are multiple publications on the survey of the iris PAD algorithms. The iris PAD methods were categorized based on the properties of iris biometric characteristic in [215] [216]. Some other surveys include a brief overview of iris PADs by Wei et al. [217] and textured contact lens PAD survey by Bowyer and Doyle

Table 9.1: Texture analysis based iris presentation attack detection algorithms.

Author	Type of Attack	Feature extraction algorithm	Dataset
Kohli *et al.* [205]	Print, Contact lens and Synthetic	Multi-order dense Zernike moments and Local Binary Patterns (LBP) with variance	IIITD
Hu *et al.* [206]	Print and Contact lens	Regional features via spatial pyramid and relational measure	Clarkson, Warsaw, Notre Dame and MobBIOfake
Raja *et al.* [213]	Print and Display	Adaptive hybrid patterns (AHP)	Warsaw '13, ATVS-Fir, MobILive 2014, VSIA, PAVID
Gragnaniello *et al.* [201]	Contact lens	Scale-invariant image descriptor	Notre Dame and IIITD
Yadav *et al.* [3]	Contact lens	Local Binary Patterns (LBP) and Pyramid Histogram of Oriented Gradients (PHOG)	IIITD and NDCLD
He *et al.* [4]	Contact lens	Gray level co-occurrence matrices (GLCM)	SJTU iris database
Komulainen *et al.* [202]	Contact lens	Binarized Statistical Image Features (BSIF)	NDCLD'13
Raghavendra *et al.* [203]	Contact lens	Statistically independent filters for Binarized Statistical Image Features (BSIF)	IIITD and NDCLD'12
Doyle and Bowyer [204]	Contact lens	Binarized Statistical Image Features (BSIF)	NDCLD'15
Raghavendra *et al.* [219]	Contact lens	Deep CNN	NDCLD'13 and IIITD
Mandalapu *et al.* [220]	Contact lens	Blind Reference less Image Quality Evaluator (BRISQUE) and BSIF	Clarkson, Notre Dame and IIITD

[218]. In the recent work, an assessment of State of the Art PAD for iris recognition by Czajka and Bowyer [28] and in [46] authors have presented the summary of popular attacks and taxonomy of PAD methods.

In the literature, a number of PAD methods on iris biometrics are proposed. Among texture-based features, Local Binary Patterns (LBP) are popularly used in PAD. A wide variety of LBPs has been used to improve the performance of the PAD. Boosted Local binary patterns (LBP) are used on sub-divided images of the normalized iris for iris PAD by He *et al.* in [208] and weighted LBPs by by Zhang *et al.* in [209]. Yadav *et al.* proposed two methods based on LBP 1)modified LBP and 2)feature-level concatenation of local binary patterns and pyramid of the histogram of oriented gradients (LBP + PHOG) in [3].

Gray level co-occurrence matrices are other texture-based features used by He *et al.* for statistical texture analysis based approach to detect artefact iris samples [4]. Scale-invariant image descriptors on segmented iris images were used by Gragnaniello *et al.* to detect contact lenses [201]. A novel texture pattern called Hierarchical Visual Codebook (HVC) was proposed by Sun *et al.* for sparse representation of the iris texture to classify artefact and bona fide iris images [210]. Raja *et al.* [213] proposed color adaptive quantized hybrid patterns for texture feature extraction to identify print and display attacks on ocular images.

Recent works in this direction focus on detecting contact lens using Binarized Statistical Image features (BSIF). In Komulainen *et al.*, BSIF features are applied from pre-processed iris images to obtain more generalized results in detecting textured contact lens [202]. Statistically independent filters were used by Raghavendra *et al.* [203] to extract BSIF features. Doyle and Bowyer [204] used BSIF filters at

multiple scales for detecting textured contact lenses. A blind referenceless image quality based feature and BSIF features are used to detect contact lens in iris images by Mandalapu *et al.* in [220].

Advanced methods use deep convolutional neural networks for improved contact lens detection [219]. Multi-spectral sensors are developed in [221] for additional iris information in multiple bands and are used to improve the iris verification and more accurate PAD. A brief overview of the state-of-the-art texture feature based contact lens presentation attack detection methods is presented in Table 9.1.

9.4 Evaluation Methodology

In this work, five presentation attack detection algorithms are selected for evaluation. Four databases consisting of bona fide and 2 different kinds of presentation attacks from different sensors are taken and experimented in 4 different scenarios. The five PAD algorithms used in this work are discussed in the following subsection.

9.4.1 Presentation Attack Detection Methods

Method 1: LBP + SVM

The Local Binary Patterns (LBP) is popular descriptor used for texture classification [222]. In this work, LBP features are used to train a support vector machine (SVM) using a linear kernel. The training support vector machines are used to compute the probability scores on the test samples. These scores are used to decide whether the test sample is a bona fide or presentation attack sample.

Method 2: mBSIF + SVM

Multiscale Binarized Statistical Image Features (mBSIF) [42] is used to explore both periocular (or eye region) and iris region to perform a presentation attack detection. The Multi-Scale Binarized Statistical Image Feature Extraction is widely used as a feasible alternative to the manual design filters as in Local Binary Patterns (LBP). The use of M-BSIF filters is proved to exhibit the characteristics like a generalization, statistical independence and robustness [42]. Four support vector machines are trained on BSIF features: 3 machines each on each feature vector from 3 filters and one on the concatenated feature vector of three features. The kernel used in support vector machines is linear. The final score of a sample is obtained by performing weighted scoring as presented.

Method 3: CAQP + SRKDA

This method is based on the framework on hybrid texture feature extraction technique that can be used for different ocular imaging systems in both NIR and visible

spectrum imaging systems [213]. This method uses adaptive and quantized hybrid texture patterns (AHP) obtained from local microfeatures and global spatial features for different color channels in an image. AHPs are designed to obtain unique features employing the concepts of angular quantization which is the main difference when compared to LBP. This method employs spectral regression kernel-based discriminant analysis to decide real or artifact iris sample [43].

Method 4: LBP + PHOG + SVM

In this method, a feature-level fusion of two texture features is performed as mentioned in [3]. The first feature is local binary patterns which are discussed earlier and the second feature is a pyramid of the histogram of oriented gradients (PHOG). PHOG is used as a local shape descriptor and the spatial layout of an image [223]. In PHOG descriptors, local shape is obtained by distribution over edge orientations within a region and spatial layout is computed by tiling the image into regions at multiple resolutions. Finally, the histogram of oriented gradients over each image sub-region at each resolution are concatenated to form PHOG. Linear SVM is used to train the concatenated feature vector of iris images.

Method 5: BRISQUE + BSIF + SRKDA

A weighted score-level fusion of two features is used in this method as proposed in [220]. The first feature is an image quality based feature called Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). BRISQUE is a natural scene statistic-based distortion-generic image quality assessment (IQA) model that operates in the spatial domain to provide the holistic measure of the image quality [41]. The Second feature is texture based feature using Binary Statistical Image Features (BSIF) using only the filters of size 7×7 with a length of 10 bits. The choice has been made by considering the performance [212] and accuracy of iris presentation attack detection from the previous works [42].

For computing the scores of the iris samples, Spectral Regression Kernel Discriminant Analysis (SRKDA) [43] is trained on the individual features and test scores are obtained. Further, a Fischer Discriminant Ratio (FDR) [44] based weighted fusion is performed on the test scores. The protocol is the same as the proposed method in [220].

9.4.2 Datasets

Four datasets from LivDet-Iris 2017 competition are used in this work. A brief description of the datasets and the example images are presented in this subsection. Table 9.2 presents types of samples and capturing devices used in each of the datasets.

Table 9.2: Description of LivDet-Iris 2017 datasets

Name	Types of samples	Capturing device
Warsaw	Live and Print	IrisGuard AD 100
Clarkson dataset	Live and Print	LG IrisAccess EOU2200 camera
Notre Dame dataset	Live and Textured Contact lens	LG 4000 and AD 100
IIITD dataset	Live and Textured contact lens	Cogent and Vista

Warsaw dataset

Warsaw dataset used in LivDet-Iris 2017 competition has been collected at the Warsaw University of Technology in Poland. It consists of 1844 images acquired for 322 distinct irises and 2669 images of the corresponding paper printouts. IrisGuard AD 100 sensor is used to capture all real and spoof samples. The liveness detection in the sensor is intentionally deactivated to acquire printouts possible. Each printout had a hole cut in a place of the pupil to generate a genuine reflection from the cornea as expected by the sensor. The resolution of samples is 640 x 480 pixels and real images are compliant with ISO/IEC 19794-6. The training and testing partitions are the same as LivDet-iris 2017. Training and testing subsets are subject-disjoint. That is, subjects selected for training subset are not present in the testing subset. The sample live and printout images of same subject are presented in Figure 9.1.

Figure 9.1: Sample images from Warsaw dataset. Left: Live, Right: Printout

Clarkson dataset

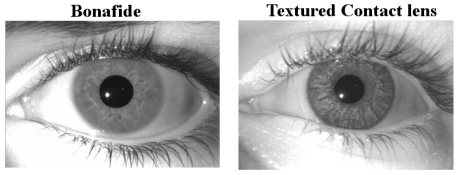
The Clarkson dataset for LivDet-Iris 2017 was collected at Clarkson University using an LG IrisAccess EOU2200 camera. The Clarkson dataset consisted of three parts. The first part is live iris images collected from cooperative subjects. The second is patterned contact lenses and the third part is printed iris images. In this work, only live and print samples were taken to evaluation. Attack images were printouts of NIR iris images of the eye. The training set consisted of 2469 live images from 25 subjects and 1346 printed images from 13 subjects. Each image

is 640 x 480 pixels. The testing set consists of 1485 live images from 25 subjects. There are 908 printed images, 764 standards printed from 12 subjects as well as 144 visible light iris images from 24 subjects. Figure 9.2 shows sample images from Clarkson dataset.

Figure 9.2: Sample images from Clarkson dataset. Left: Live, Right: Printout

Notre Dame Contact Lens dataset

The Notre Dame contact lens dataset was captured using two sensors LG 4000 and AD 100. The image resolution of all samples is 640 x 480 pixels. Both the training and testing set of LG 4000 sensor data contains 1000 live images and 400 textured contact lens each. The AD 100 sensor data contains 200 live images and 100 textured contact lens images in each of the training and testing tests. All real samples are compliant with ISO/IEC 19794-6 and only texture contact lens was used. The soft (or transparent) contact lens was excluded from the data similar to the procedure in LivDet-Iris 2017. The textured contact lens used in this dataset were from five different manufacturers: J&J, Ciba, Cooper, UCL and ClearLab. Figure 9.3 shows the example images of live and textured contact lens from two sensors in Notre Dame dataset.



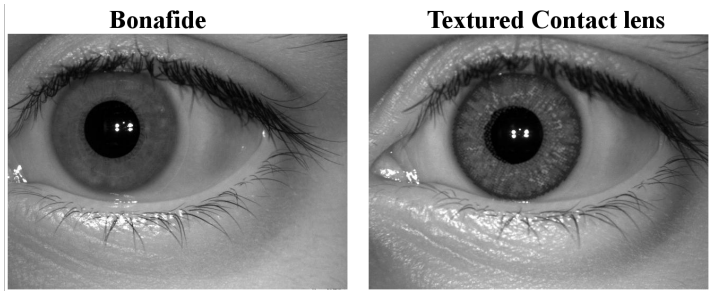
(a) LG 4000 sensor

(b) AD 100 sensor

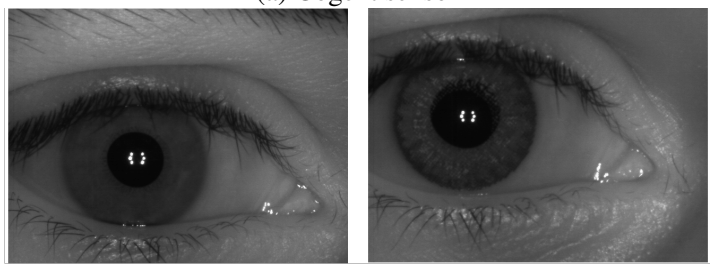
Figure 9.3: Example images of bona fide samples and textured contact lens attack samples from two sensors of the Notre Dame dataset

IIITD Contact Lens Iris (CLI) dataset

The IIITD contact lens dataset also contains textured contact lenses from various manufacturers: CIBA Vision Freshlook Dailies, Bausch and Lomb Lacelle, and Aryan 3-Tone. Two sensors: Cogent and Vista were used in capturing the images. The IIITD Cogent dataset contains 1143 live and 1138 textured contact lens images. 700 bona fide images and 700 textured contact lens images of the Cogent dataset were taken for training, development and rest (443 bona fide and 438 textured contact lens) are used for testing. The IIITD vista dataset contains 990 bona fide and 990 textured contact lens images. The vista dataset is divided into 500 bona fide images and 500 textured contact lens images for training, development and rest of the samples (490 bona fide and 490 textured contact lens) are taken for testing. The sample images from IIITD Contact lens dataset from two sensors are shown in Figure 9.4.



(a) Cogent sensor



(b) Vista sensor

Figure 9.4: Example images of bona fide samples and textured contact lens attack samples from two sensors of the IIITD dataset

9.4.3 Performance protocol

In this work, we present the results by following the International Standard ISO/IEC 30107-3 [37] in terms of Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER). APCER and BPCER metrics are described as follows:

- Attack Presentation Classification Error Rate (APCER) is the proportion of attack presentations that are incorrectly classified as bona fide presentations.
- Bona fide Presentation Classification Error Rate (BPCER) is the proportion of bona fide presentations that are incorrectly classified as presentation attacks.

We computed two BPCER values for each dataset, B_5 and B_10, by fixing APCER at 5% and 10% respectively. Also, we include the results in terms of Detection Equal Error Rate (D-EER%) and detection error trade-off curves are presented for each case.

9.5 Experiments and Results

Four different types of scenarios are identified and experiments are designed to evaluate the performances of different PAD methods towards two different kinds of attacks on two different datasets each.

9.5.1 Experiment 1: Individual PAD evaluation

In this experiment, the detection performance of each PAD algorithm is evaluated individually on each dataset. Two types of attacks are evaluated separately and results are presented. The performance of print attack on Clarkson and Warsaw dataset are compared in Table 9.3 and textured contact lens attack detection accuracy is presented in Table 9.4.

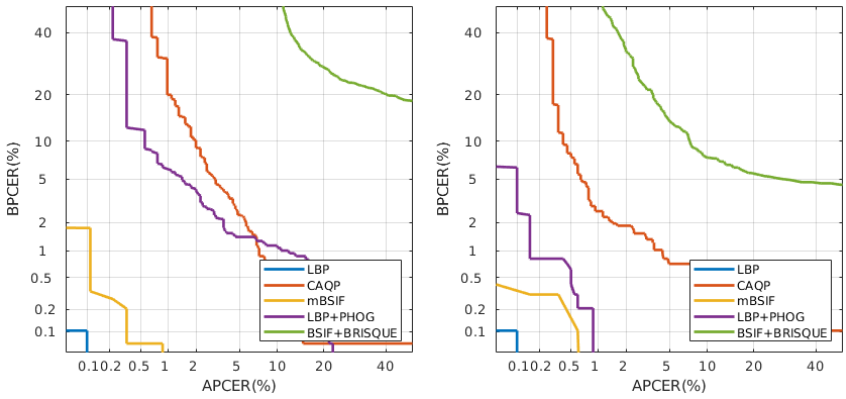
Case 1: Print-attack detection

Table 9.3: Performance of Print attack detection

	Clarkson Dataset			Warsaw Dataset		
	D-EER	**B_5**	**B_10**	**D-EER**	**B_5**	**B_10**
LBP	0.1009	0	0	0.1007	0	0
CAQP	3.85	2.36	0.33	1.84	0.71	0.41
mBSIF	0.24	0	0	0.30	0	0
LBP + PHOG	2.72	1.41	1.14	0.50	0	0
BSIF + BRISQUE	24.43	94.33	66.53	8.6	13.75	7.49

The important observations made from this experiment are as follows.

1. The Local Binary Patterns (LBP) has performed with near zero error on print



(a) Clarkson dataset (b) Warsaw dataset

Figure 9.5: DET curves: PAD evaluation of Print attack

attacks on two different datasets.

2. Other PAD methods also display very low error rates which proves that the problem of print attacks can be easily solved when tested samples are similar to the ones the PAD method is trained on.

3. Hybrid methods like feature-level and score-level fusion algorithms (Method 3, 4 and 5) did not improve the performance of the PAD method.

4. The image quality and texture feature-based method (Method 5) display the least accuracy in detection print attacks even though it performs better on contact lens attacks [220].

5. The comparison of the error rates are presented in the Detection error tradeoff (DET) curves of the five algorithms for two print attack datasets in Figure 9.5.

Case 2: Contact lens attack detection

The segmented and normalized iris is used to perform contact lens detection. The results of the experiment are presented in Table 9.4

The key points observed from this experiment are mentioned below.

1. The mBSIF (Method 3) display the highest average accuracy on both datasets in detecting textured contact lens with a D-EER of 2.12%.

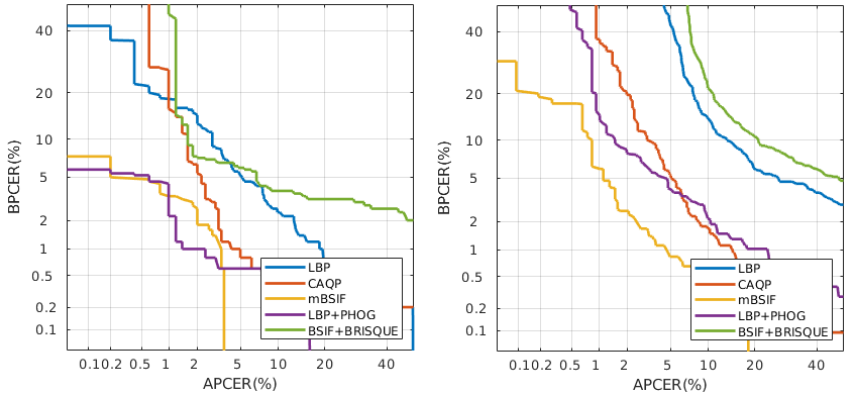
Table 9.4: Performance of Contact-Lens attack detection

	Notre Dame Dataset			IIITD Dataset		
	D-EER	**B_5**	**B_10**	**D-EER**	**B_5**	**B_10**
LBP	5.2	5.6	2.4	11.34	44.10	14.04
CAQP	3	1	0.6	5.06	5.43	1.68
mBSIF	2	0	0	2.25	0.84	0.18
LBP + PHOG	1.2	0.6	0.4	4.77	4.02	2.05
BSIF + BRISQUE	5.8	6	3.8	14.33	78.74	21.34

2. Though LBP + PHOG (Method 4) show low error rates on Notre Dame datasets, the error rates have increased in case of IIITD dataset. This might support the statement in [3] that sample quality of IIITD dataset is lower than Notre Dame dataset.

3. Using multiple texture filters based fusion method (Method 3 and 4) appears to improves the overall performance of textured contact lens detection.

4. The Detection error tradeoff (DET) curves of the five algorithms for two contact lens attack datasets are presented for comparison in Figure 9.6.



(a) Notre Dame dataset (b) IIITD dataset

Figure 9.6: DET curves: PAD evaluation of Contact Lens attack.

9.5.2 Experiment 2: Cross-dataset evaluation

In this experiment, the cross-dataset evaluation of PAD methods is evaluated. For this, one dataset is used for developing the PAD system and other dataset is used for testing. Table 9.5 and 9.6 contains the error rates of cross-dataset evaluation.

Case 1: Print Attacks

Table 9.5: Cross-dataset Evaluation of Print attack

Train	**Clarkson dataset**			**Warsaw dataset**		
Test	**Warsaw dataset**			**Clarkson dataset**		
	D-EER	**B_5**	**B_10**	**D-EER**	**B_5**	**B_10**
LBP	33.95	72.79	69.19	41.20	99.59	99.39
CAQP	46.5	94.35	90.45	35.68	70.04	59.51
mBSIF	6.05	6.26	4.51	18.19	41.7	27.8
LBP + PHOG	32.08	68.89	61.70	37.69	99.52	97.36
BSIF + BRISQUE	2.44	0.92	0.82	43.32	92.10	85.96

The main observations from the cross-dataset evaluation of PAD methods on print attack detection are as follows.

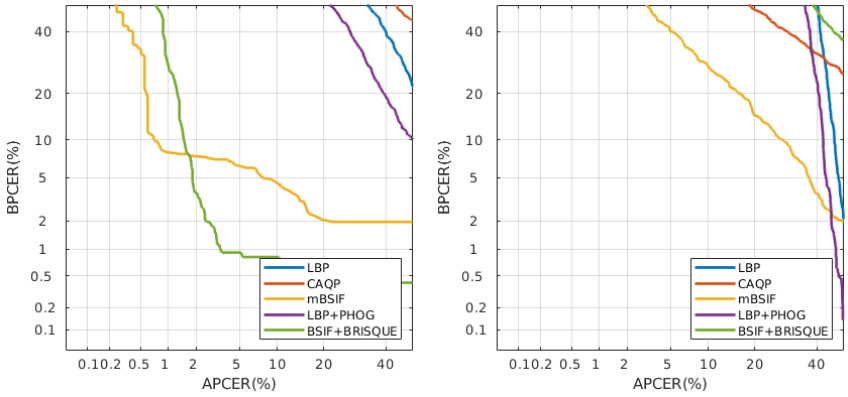
1. The best performance is observed in the case of mBSIF method (Method 3) with an average D-EER of 12.12%.

2. Method 5 shows the lowest error rates when the PAD system is developed using Clarkson dataset and tested on Warsaw dataset. Similarly, Method 2 shows the lowest error when using Warsaw dataset for training and tested on Clarkson dataset. However, in both cases, the error rates raised abruptly considering other cases. This makes the generalization properties of PAD methods highly questionable.

3. The comparison of APCER and BPCER error rates of cross-dataset evaluation in case of print attacks is presented through Detection error tradeoff (DET) curves in Figure 9.7.

Case 2: Contact Lens Attacks

The main observations from the cross-dataset evaluation of PAD methods on contact lens attack detection are as follows.



(a) Train:Clarskon, Test:Warsaw (b) Train:Warsaw, Test:Clarkson

Figure 9.7: DET curves: Cross-dataset evaluation of Print attack datasets

Table 9.6: Cross-dataset Evaluation of Contact-Lens attack

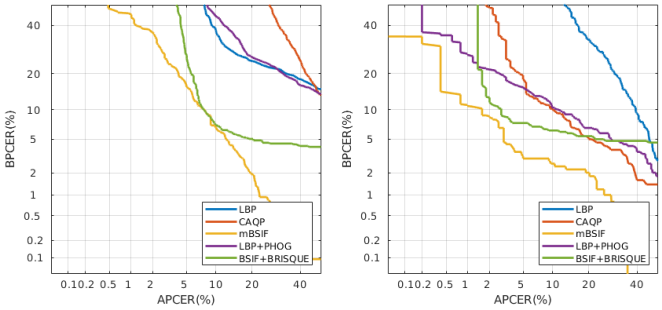
Train	**Noter Dame dataset**			**IIITD dataset**		
Test	**IIITD dataset**			**Notre Dame dataset**		
	D-EER	**B_5**	**B_10**	**D-EER**	**B_5**	**B_10**
LBP	23.24	86.6	37.54	26	63.2	54.6
CAQP	33.78	96.62	90.54	10	19.6	10.2
mBSIF	8.8	17.04	6.64	4	3.2	2.6
LBP + PHOG	24.18	68.8	44.47	10.4	15.8	10.4
BSIF + BRISQUE	8.66	27.80	7.11	6.8	7.4	6.2

1. Method 3 (mBSIF) perform better in the cross-dataset evaluation of contact lens detection with error rates lower than 10%. This again proves that using multiple texture filters improves the performance of contact lens detection even in the cross-dataset evaluation.

2. Method 5 (BSIF + BRISQUE) shows a performance closer to the best performing method. This can be attributed to usage of the fused method of image quality and texture feature can improve the robustness of PAD method to cross-dataset of evaluation.

3. Other methods show error rates more than 3 times the above two methods. By this, we can say cross-dataset evaluation in case of contact lens detection is a significant problem to consider.

4. The Detection error tradeoff (DET) curves of the cross-dataset evaluation of the five algorithms on contact lens attack datasets are compared in Figure 9.8.



(a) Train:Notre Dame, Test:IIITD (b) Train:IIITD, Test:Notre Dame

Figure 9.8: DET Curves: Cross-dataset evaluation of Contact lens datasets

9.5.3 Experiment 3: Unknown attack detection

This experiment focuses on the scenario where the type of attacks is not known in advance. That is, the PAD methods use one kind of attacks where the target samples contain different attacks. For example, the training samples are only from live and print attacks but the testing samples include live and contact lens attack samples. This is a common situation in real-world application due to the growth of new kind of presentation attacks which are unknown to the PAD development system. This experiment examines the generalization abilities of the PAD algorithms. The results of this experiment are presented in Table 9.7.

Table 9.7: Cross-dataset Evaluation of Print attack

Train	**Print**			**Contact lens**		
Test	**Contact lens**			**Print**		
	D-EER	**B_5**	**B_10**	**D-EER**	**B_5**	**B_10**
LBP	50	98.46	97.70	50	99.8	99
CAQP	48.72	95.08	90.88	31.39	74.71	61.64
mBSIF	41.19	90.11	83.03	22.30	63.27	45.19
LBP + PHOG	50	95.91	90.56	50	97.55	91.16
BSIF + BRISQUE	25.94	92.66	74.74	6.6	23.94	1.5

The main observations from the unknown attack detection evaluation of PAD methods are as follows.

1. Method 5 (BSIF + BRISQUE) based on image quality and texture features performed better when trained with print and tested on contact lens data and the other way. This display the robustness of the method to new kind of attacks.

2. All other methods show very high error rates in both cases of print and contact lens detection. This indicates that unknown attacks are a challenging problem.

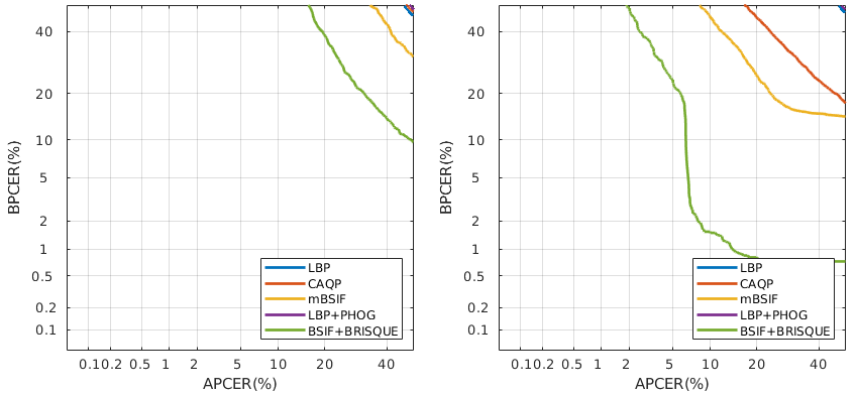
3. The Detection error tradeoff (DET) curves of unknown attack detection also indicates the deficient performance of the PAD methods as shown in Figure 9.9.

9.5.4 Experiment 4: Multi-Attack Multi-Sensor scenario

In Experiment 4, we considered a scenario where samples from multiple attacks and multiple sensors are involved in the development of PAD systems. Therefore, we combined the training data from all four datasets and two attacks and prepared PAD methods. Similarly, for testing, we combined all testing samples and performance score is computed. Table 9.8 shows the results obtained in this experiment.

The main observations from the Multi-Attack Multi-Sensor evaluation of PAD methods are as follows.

1. The CAQP method show very low errors in this experiment with D-EER of



(a) Train:Print, Test:Contact Lens (b) Train:Contact lens, Test:Print

Figure 9.9: DET Curves: Unknown attack evaluation

Table 9.8: Multi Attack Multi Sensor evaluation

	D-EER	**B_5**	**B_10**
LBP	7.64	11.90	5.19
CAQP	2.83	1.01	0.47
mBSIF	13.67	32.03	18.86
LBP + PHOG	3.00	1.368	0.86
BSIF + BRISQUE	16.8	84.66	44.28

2.83%.

2. Also LBP + PHOG method display performance very close to CAQP method with D-EER of 3%. Thus, the usage of multiple-attacks and multi-sensors for implementing a PAD method with multiple texture filters can improve the performance of accuracy of presentation attack detection.

3. Other methods also shows error rates not higher than 15% when using data from all types of attacks and all sensors to train PAD method.

4. It is also observed in Figure 9.10 that DET curves of best methods are displayed a similar change in error rates.

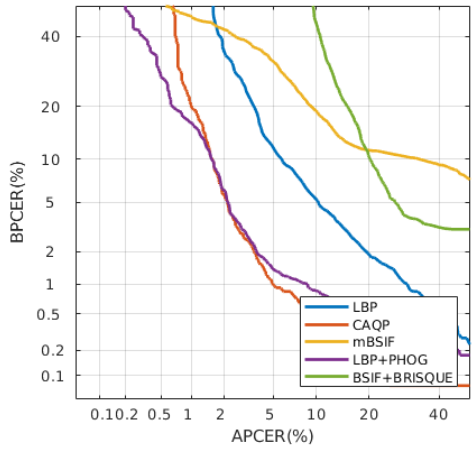


Figure 9.10: DET curves: Multi-Attack Multi-Sensor PAD evaluation

9.6 Conclusion

The human iris shows better properties of uniqueness when compared to other biometric characterisitics (face and fingerprint). Several high-security applications use Iris-based biometric identification. The vulnerabilities in iris recognition systems make their use highly questionable. More particularly, presentation attacks cause severe threats to iris biometric systems. Due to the high availability of different kinds of artefacts and rising growth in technology, preparing presentation attack instruments has becoming simple. This makes the development of high performing presentation attack detection algorithms quite challenging.

In this work, a survey is performed on texture-based state-of-the-art presentation attack detection (PAD) algorithms on iris. The survey presents the types of methods used in individual presentation attacks and discussions are provided. Four dif-

ferent publicly available datasets from LivDet Iris 2017 namely, Warsaw, Clarkson, Notre Dame and IIITD, are chosen for the evaluation of two important presentation attacks namely, prints and contact lens. We choose 5 different texture-based state-of-the-art PAD methods to evaluate and compare the performances. We followed ISO/IEC performance evaluation protocols [37] to obtain error rates on 5 datasets in 4 different scenarios of presentation attacks.

Table 9.9: Results of all experiments and evaluations

Type of Experiment	**Type of Attack**	**Best performed PAD method**
Experiment 1 PAD Evaluation	Print	LBP
	Contact Lens	mBSIF
Experiment 2 Cross-dataset	Print	mBSIF
	Contact lens	
Experiment 3 Unknown attacks	Print-Contact lens	BSIF+BRISQUE
	Contact lens-Print	
Experiment 4 Multi-attack Multi-sensor	All combined	CAQP

The experimental results and comparisons of the PAD algorithms are presented in detail. It is clearly observed that the performance of PAD algorithms vary depending on the scenario of presentation attacks. In Table 9.9, best performed PAD algorithm in each scenario is presented. For direct print attack, a simple textured based approach using local binary patterns is performing better than other methods with near-zero error rates. In the case of contact lens detection, the mBSIF method shows very high accuracy. In the scenario of cross-dataset evaluation, mBSIF method outperformed other methods in both print and contact lens attack detection.

The scenario of unknown attacks is a challenging problem due to the evolving new kinds of presentation attacks. The texture and image quality fusion method of BSIF+BRISQUE shows high performance with error rates close to half of the next best performing method. In the scenario of using data from all attacks and sensors in developing PAD method, color adaptive quantized patterns(CAQP) method display high accuracy in detecting the attacks. Though no single method performs better in all scenarios, the results show that the fusion of multiple high performing texture features will deliver higher accuracy in detecting presentation attacks.

Chapter 10

Article 5: Multilingual voice impersonation dataset and evaluation

Mandalapu, Hareesh, Raghavendra Ramachandra, and Christoph Busch. "Multilingual voice impersonation dataset and evaluation." In *International Conference on Intelligent Technologies and Applications*, pp. 179-188. Springer, Cham, 2020.

10.1 Abstract

Well-known vulnerabilities of voice-based biometrics are impersonation, replay attacks, artificial signals/speech synthesis, and voice conversion. Among these, voice impersonation is the obvious and simplest way of attack that can be performed. Though voice impersonation by amateurs is considered not a severe threat to ASV systems, studies show that professional impersonators can successfully influence the performance of the voice-based biometrics system. In this work, we have created a novel voice impersonation attack dataset and studied the impact of voice impersonation on automatic speaker verification systems. The dataset consisting of celebrity speeches from 3 different languages, and their impersonations are acquired from YouTube. The vulnerability of speaker verification is observed among all three languages on both the classical i-vector based method and the deep neural network-based x-vector method.

10.2 Introduction

Biometric authentication for providing access to information, devices, and networks have been used in security applications for many years. Speaker recognition

is one of the modalities that has been prominently used as biometrics for the last few decades. Though computational intelligence has advanced, biometric systems are still vulnerable in the authentication of individuals. In voice-based person verification, there are emerging new ways of attacks every day. The popular speaker verification vulnerabilities are voice impersonation, audio replay attack, speech synthesis, and voice conversion. Though speech synthesis and voice conversion can cause severe impact, these attacks can only be performed with certain access to the biometric system. The conventional physical access attacks can only be performed by voice impersonation or replay attacks.

Voice impersonation is discussed to be having minimal impact on speaker recognition systems when compared to other kinds of attacks [224]. However, studies have shown that a professional impersonator having enough training on the target's speech can perform a successful attack [225][226]. It is also a simple way of attacking a voice-based biometric system. By adjusting the vocal cords, an impersonator can mimic a target speaker's voice. Though it has been observed that it is difficult to impersonate untrained target's voices, well-known impersonators after multiple attempts can successfully attack a speaker recognition system. Automatic speaker verification and the vulnerability evaluation have multiple dependencies like text, language, and channel effects [227]. After considering the issues mentioned above, there is a requirement of research work in fully understanding the effect of impersonation with all the dependencies.

In this work, two popular speaker recognition systems are evaluated over the effect of impersonation. We have included three different languages with no text-dependency and various channel data to accommodate the previously mentioned dependencies of automatic speaker verification. Further, this work is organized as follows. A literature review on the previous studies on voice impersonation is presented in Section 10.3. In Section 10.4, the impersonation dataset captured is mentioned with details. The Automatic speaker Verification methods chosen for our experiments and trained dataset used are discussed in Section 10.5. Section 10.6 explains the experiments performed and results obtained in impersonation vulnerability evaluation. The conclusion of this work and future directions are presented in Section 10.7.

10.3 Related Work

In the initial works, amateur impersonators were used in performing attacks. Lau et al. [228] have performed experiments on the YOHO dataset, which contains 138 speakers. Two subjects acted as impersonators, and the vulnerability of the speaker recognition system towards such mimicry attack was verified. Upon performing multiple attempts, it was observed that an impostor could perform an at-

tack if the impostor has the knowledge about enrolled speakers in the database [229]. In [230], Mariéthoz et al. assessed the vulnerability of state-of-the-art text-independent speaker verification system based on Gaussian mixture models (GMMs) to attacks conducted by a professional imitator. It was observed that the GMM based systems are robust to mimicry attacks.

Farrús et al. [225] performed experiments on prosodic features extracted from voices of professional impersonators to perform mimicry attacks on speaker identification systems. The increase in acceptance rates was observed when imitated voices are used for testing. Panjwani et al. [231] proposed a generic method and used crowd-sourcing for identifying impersonators. The GMM-UBM based method displays an increase in impostor attack presentation match rate (IAPMR) when using professional impersonators. Hautamäki et al. [226] used three modern speaker identification systems to test the case of voice mimicry. It has been observed that the EER values for GMM-UBM based method are decreased but increased for two other i-vector based methods.

The ASVspoof (Automatic Speaker Verification spoof) challenges are a series of evaluations focus on improving countermeasures to attacks on speaker verification systems. Voice conversion and speech synthesis attacks are the primary focus in the first ASVspoof challenge [224]. The Second ASVspoof challenge is evaluated for countermeasures to different kinds of replay attacks [232]. The recent challenge in this series includes both physical (replay attacks) and logical access (voice conversion, speech synthesis) attacks [40]. Impersonation attacks are not considered in any series of these competitions, mentioning that impersonation's relative severity is uncertain. However, the attacks discussed in these series assumed to have access to the biometric system. For example, the audio sample's digital copy is necessary to perform replay attacks, and logical access attacks need access into the system where the digitally manufactured copy of utterance is presented. Impersonation is a physical access attack on voice-based biometrics that does not require any access to the biometric system, which makes it an interesting research topic for this study.

It was observed in most of these methods that voice impersonation has a considerable impact on speaker verification systems, but all these methods possess certain challenges, which are observed as follows.

- There is no publicly available impersonation attack dataset similar to other attacks like replay, voice conversion, and speech synthesis. Also, there is a requirement of professional impersonators to compose a dataset.

- State-of-the-art speaker verification systems are not employed in the evalu-

ation.

- The text-dependent methods are used to perform an attack, which is not a generalized scenario.
- The impact of language and channels are not discussed in the previous evaluations.
- Standard protocols were not used to evaluate the impact of impersonation.

The following contributions are made in this paper to address the challenges mentioned above.

- A dataset of bona fide and impersonator samples is created from YouTube videos for three different languages, which will be made publicly available (similar to VoxCeleb dataset).
- Three different languages, text-independent speeches, and multiple channel data are captured in the dataset.
- Extensive experiments are carried out on one classical and one state-of-the-art speaker verification systems in three different languages.
- Results are presented following ISO/IEC standards for biometric system performance evaluation and presentation attack detection.

10.4 Voice Impersonation Dataset

The dataset of bona fide speeches and corresponding impersonated speeches are acquired in a process similar to that of the VoxCeleb database. The easiest way to obtain this type of attack dataset is by looking for popular people and their impersonators' speeches that are uploaded to YouTube. In this work, three languages are chosen as per the authors' knowledge: English and two Indian languages: Hindi and Telugu. Multi-lingual data samples also help us to understand the impact of language used in training data on ASV systems. The bona fide speakers and their well-known impersonators are carefully selected from different subjects in each language. The speakers include political figures and actors.

The bona fide speeches are taken from the interview videos of the target speakers. The impersonation speeches are obtained from YouTube videos of television shows and performances by mimicry artists ranging from amateurs to professionals. The speeches are manually annotated and segmented to individual speakers without any loss in the quality of audio. The speech samples with dominating

Table 10.1: Details of impersonation attack dataset.

Language	No. of speakers	Bona fide utterances	Impersonation utterances
English	15	506	411
Hindi	15	768	449
Telugu	15	677	549

Table 10.2: Details of the verification split of VoxCeleb1 dataset

VoxCeleb1	Dev Set	Test Set
No.of Speakers	1211	40
No.of Videos	21,819	677
No.of Utterances	148,642	4,874

background noise like applause and music are ignored. The number of speakers and utterances for each language in this dataset is presented in Table 10.1.

10.5 Vulnerability of ASV systems to Voice Impersonation

The impact of voice impersonation on automatic speaker verification (ASV) systems are verified by performing a presentation attack on the ASV methods using impersonation samples. The initial step in this process is to acquire voice impersonation samples for a set of speakers. Due to the lack of professional impersonators for several speakers, and based on the authors' knowledge of target speakers, we have chosen an obvious way of obtaining impersonation samples from YouTube and included three different languages.

10.5.1 Training Dataset

In our work, we have used the pre-trained models [1] from Kaldi toolkit [63]. The models are trained on verification split of the VoxCeleb1 and entire VoxCeleb2 dataset [233]. The training dataset is a part of the VoxCeleb dataset, which is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. The main reasons for choosing VoxCeleb trained model are a huge variety of speakers and samples in the dataset (more than 1 million samples and over 7200 speakers) and also the similarity to our dataset of mimicry samples from YouTube. The training dataset contains speech from speakers of a wide variety of cultures, accents, professions, and ages. The details of dataset is presented in Table 10.2 and 10.3.

[1] VoxCeleb Models: http://kaldi-asr.org/models/m7

Table 10.3: Details of VoxCeleb2 dataset

VoxCeleb2	Dev Set	Test Set
No.of Speakers	5994	118
No.of Videos	145,569	4911
No.of Utterances	1,092,009	36,237

10.5.2 Automatic Speaker Verification (ASV) Systems

The next step is to obtain ASV systems to examine vulnerability due to voice impersonation. We chose two different methods for this purpose 1. a classical i-Vector based system and 2. a state-of-the-art deep neural network-based x-vector method.

I-vector Method

The I-vector based automatic speaker verification method is the state-of-the-art approach proposed in [188]. I-vectors are the low dimensional representation of a speaker sample that is estimated using Joint Factor Analysis (JFA), which models not only the channel effects but also information about speakers. With the help of i-vector extraction, a given speech utterance can be represented by a vector, which includes total factors. The channel compensation in i-vectors is carried out in a low-dimensional total variability space. In this method, we have employed probabilistic linear discriminant analysis (PLDA) [31] to train the speaker models. The trained PLDA models are then used to compute the log-likelihood scores of the target samples to verify the speaker.

X-Vector Method

The deep learning and end-to-end speaker verification approaches are the recent popular methods replacing handcrafted methods. The x-vector based speaker verification is one of the latest approaches using deep neural network (DNN) embeddings [32]. This approach uses trained DNN to differentiate speakers by mapping their variable-length utterances to a fixed-dimensional embedding called as x-vectors. A large amount of training data is one of the biggest challenges in this approach. Therefore, data augmentation with added noise and reverberation is used to increase the size of training data.

In the implementation of ASV methods, we have used the pre-trained Universal Background Models, i-vector extractor, x-vector extractor, and speaker recognition codes from Kaldi [2].

[2]Kaldi GitHub: https://github.com/kaldi-asr/kaldi

Table 10.4: Performance of ASV methods on VoxCeleb1 test set

ASV method	EER (%)
i-vector method	5.3
x-vector method	3.1

Table 10.5: Equal Error Rate (EER%) values of zero-effort impostors and impersonation attacks for the ASV methods on each language

Language	Scenario	i-vector method	x-vector method
English	zero-effort impostors	5.99	3.83
	impersonation attacks	12.94	11.10
Hindi	zero-effort impostors	7.88	5.72
	impersonation attacks	11.17	12.22
Telugu	zero-effort impostors	4.84	3.86
	impersonation attacks	5.57	4.77

10.6 Experimental Results and Discussion

The test set of the VoxCeleb1 dataset is used to verify the performance of obtained ASV methods using pre-trained models. The results of ASV methods on the VoxCeleb1 test set are in Table 10.4. The thresholds used for attack samples matching bona fide samples are from this test set evaluation.

The performance of the speaker recognition systems is evaluated using the standardised metrics from ISO/IEC on biometric performance [234]. In addition the Equal Error Rate is reported. The Equal Error Rate (EER) is the rate at which false match rate (FMR) and false non-match rate (FNMR) are equal. The detection error trade-off (DET) curve is used to plot the relationship between the false match rate (FMR) and the false non-match rate (FNMR) for zero-effort impostors and impersonation attacks. Further, the impostor attack presentation match rate (IAPMR) is calculated for each language in two ASV methods. Impostor attack presentation match rate (IAPMR) is the proportion of impostor attack presentations using the same PAI species in which the target reference is matched [37]. In this case, it is the percentage of impersonation attack samples when matched with target speakers above the threshold, which is set by the test set for each ASV system.

10.6.1 Equal Error Rate (EER) comparison

The EERs (%) are presented in Table 10.5 for both zero-effort impostors and impersonation presentation attacks in order to compare the vulnerability caused by

voice impersonation on ASV methods. The zero-effort impostors' evaluation is performed with no targeting attacks, whereas the presentation attacks are evaluated by presenting attack samples targeting corresponding speakers. It is important to remember that the zero-effort impostor scores are computed by targeting one speaker on other speakers only in the same language. However, the impersonation samples of one speaker are intended only to target that particular speaker. The IAPMR values that are presented show how many attack samples are matched with target speakers' bona fide samples.

The results show that the increase in the EER (%) values when impersonation attacks are performed. The vulnerability due to the voice impersonation can be seen in both ASV methods. Although the x-vector method has better performance without any attacks (in zero-effort impostors), it can be seen that the vulnerability due to impersonation is similar to i-vector based method. This raises the point that impersonation attacks have an impact even on an advanced deep neural network-based approach similar to the classical method. The comparison of the impact of impersonation attack among different languages deduces some important points. It is interesting to see that the impact is high in the English language when compared to other languages. The reason for this could be that the language in the training dataset is English. This makes ASV methods to recognize the English impersonators more efficiently than other languages.

10.6.2 FMR vs FNMR comparison

The False match rates versus false non-match rate comparisons show the performance of a biometric system by examining the rate of mismatches in both bona fide and impostor samples. We have fixed the false match rate at 0.001 for each case of zero-effort impostors and attacks, then obtained thresholds to compute the false non-match rate. This shows the number of bona fide samples that are not allowed into the system with a fixed allowance of impostors into the system.

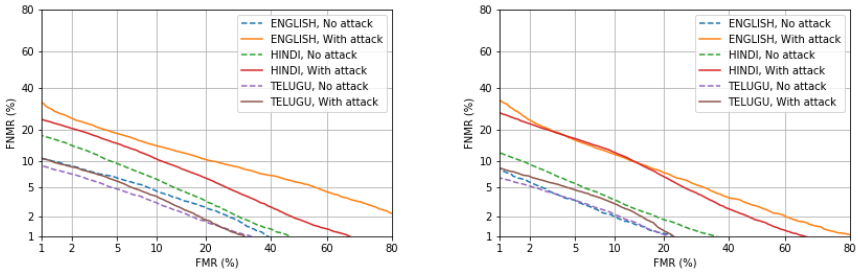
The increase in the amount of bona fide samples that result in false match is observed in all languages when attacks are performed. The highest number of mismatches can be seen in the English language in x-vector based method, where more than 66% of FNMR is observed. Further, the DET curves in Figure 10.1 shows the FMR versus FNMR of two methods in different languages with and without attacks. The increase in error rates can be seen among all systems when the impersonation attack is carried out among all three languages.

10.6.3 IAPMR evaluation

The IAPMR values in Table 10.7 show the percentage of impersonation attack samples that are matched with bona fide samples in each language. The classical i-

Table 10.6: False non-match rate (FNMR %) of zero-effort impostors and impersonation attacks when False match rate is at 0.001 (i.e. FMR = 0.1%) on each language.

Language	Scenario	i-vector method	x-vector method
English	zero-effort impostors	18.23	16.36
	impersonation attacks	51.93	66.22
Hindi	zero-effort impostors	27.43	22.63
	impersonation attacks	37.29	44.74
Telugu	zero-effort impostors	15.34	12.04
	impersonation attacks	18.31	14.55



(a) i-vector method (b) x-vector method

Figure 10.1: Detection Error Tradeoff (DET) curves of the ASV methods with and without impersonation attacks..

Table 10.7: IAPMR (%) values of the impersonation attacks.

Language	i-vector method	x-vector method
English	62.87	58.14
Hindi	46.97	53.87
Telugu	33.43	41.90

vector based method has 62.87% of attacks matched in English, which is a considerable amount showing the reasonable impact of voice impersonation on the ASV method. The state-of-the-art x-vector method accepts 58.14% of the samples. This displays a high vulnerability of the ASV method towards impersonation even on the state-of-the-art methods. For other languages Hindi and Telugu, IAPMR values are lower, which shows the language dependency of the speaker recognition. It is interesting to see that the x-vector method has a higher impact than i-vector method in Hindi and Telugu, unlike English. This impact can also be due to the dependency on the language used in training, which is English.

10.7 Conclusion

Impersonation attack have been considered as an obvious way of attacking an automatic speaker verification system. In this work, we have studied previous works on voice impersonation evaluation, and a novel dataset of voice impersonation is created. The dataset is captured in a similar way of the VoxCeleb data capturing mechanism in three different languages. The vulnerability of voice impersonation as an attack is examined on a classical and another state-of-the-art speaker recognition systems. The state-of-the-art speaker recognition method is based on a deep neural network-based method that resembles the current technology. Experiments are performed, and evaluations are carried out using ISO/IEC standards with EER, FMR/FNMR, and IAPMR metrics. The results show that the voice impersonations make the ASV methods vulnerable, with many attacks being accepted by the system. It is also interesting to see the vulnerability variation among different languages. The future works on this topic will examine the specific characteristics of the impersonator that are useful in making a successful attack on ASV methods. Also, choosing a training dataset with different languages to examine the language dependency of ASV methods and working on speaker-specific features, like residual phase, to avoid the vulnerability caused by impersonation.

Chapter 11

Article 6: Cross-lingual speaker verification: Evaluation on x-vector method

Mandalapu, Hareesh, Thomas Møller Elbo, Raghavendra Ramachandra, and Christoph Busch. "Cross-lingual speaker verification: Evaluation on X-vector method." In *International Conference on Intelligent Technologies and Applications*, pp. 215-226. Springer, Cham, 2020.

11.1 Abstract

Automatic Speaker Verification (ASV) systems accuracy is based on the spoken language used in training and enrolling speakers. Language dependency makes voice-based security systems less robust and generalizable to a wide range of applications. In this work, a study on language dependency of a speaker verification system and experiments are performed to benchmark the robustness of the x-vector based techniques to language dependency. Experiments are carried out on a smartphone multi-lingual dataset with 50 subjects containing utterances in four different languages captured in five sessions. We have used two world training datasets, one with only one language and one with multiple languages. Results show that performance is degraded when there is a language mismatch in enrolling and testing. Further, our experimental results indicate that the performance degradation depends on the language present in the word training data.

11.2 Introduction

Biometrics characteristics are used to recognize or verify the identity of a person and to provide access to the security sensitive applications. The biometric characteristics are of two different kinds: physical and behavioral. Face, fingerprint, iris are popular physical characteristics that have been in research for many years. Behavioral biometrics are based on the way humans perform certain tasks like speaking and walking. Speaking characteristics of humans are a well-known biometric modality used to perform accurate recognition. Automatic Speaker Verification has been a famous topic in biometric applications for many years now.

The advancement of computational abilities in the recent decades encouraged applications to use biometric algorithms in many fields. Due to he wide variety of users, devices, and applications, many kinds of vulnerabilities and dependencies are evolved in operational biometric systems. The popular vulnerabilities are anomalies in the samples and presentation attacks on the biometric devices. The dependencies are caused due to data capturing methods, change in devices, aging of the subject, and many more. There are more dependencies on behavioral biometric modalities because the behavior of the subject changes often. In speaker recognition, apart from the capturing conditions like microphone and transmission channel, background noise, the biometric algorithms also depend on the text, language, and emotion which impact the voice sample [48].

Text-dependent speaker recognition has been in use for many years [235]. In these kinds of approaches, the set of words used in testing is a subset of the words used in enrolment. Further, text-independent speaker recognition methods using Gaussian mixture models are introduced [236], and more algorithms were proposed to exclude the dependency caused by the text [227]. Language dependency is another challenging problem that emerged due to multilingual subjects and wide usage of the same biometric algorithm across the world. Language-independent approaches have been proposed on top of text-independent speaker recognition methods [49] by including multiple languages in enrolment. The National Institute of Standards and Technology Speaker Recognition Evaluation (SRE) series has been including multiple languages in their evaluation protocols over the years [1].

In this work, cross-lingual speaker verification is evaluated on a smartphone based dataset with different languages. The objective is to benchmark the performance of the state-of-the-art algorithms when different languages are mismatched in training, enrolling, and testing phases of automatic speaker verification. Thus, the following are the main contributions of this paper:

[1] https://www.nist.gov/itl/iad/mig/speaker-recognition

- Experiments on state-of-the-art methods that use advanced deep neural networks, like x-vector method, to check the language dependency.
- Experiments on multiple languages and multiple session datasets are included in this work.
- The dependency of trained languages used in world training data is evaluated.
- Results and discussions are presented using ISO/IEC standardized metrics for biometric performance [234].

The rest of the paper is organised as follows: Section 11.2.1 discusses the previous works on cross-lingual speaker recognition approaches and challenges. Section 11.3 describes the state-of-the-art approaches chosen for our experiments. In Section 11.4, the multilingual dataset is described, and Section 11.5, the cross-lingual experiments are presented with results and discussed. Finally, Section 11.6 concludes the work with the presentation of future work.

11.2.1 Related Work

The Automatic speaker verification as a biometric modality has emerged into many applications. The initial problems in speaker recognition have leaned over the text-dependency of the speeches in different speaker verification modules. Later, the language dependency has emerged into a challenging problem in text-independent speaker verification [49]. The early works on language mismatch evaluation are performed by comparing speaker verification with world models trained on only one language and multiple languages. One could observe that when provided with all languages and enough data for world model training, there is no degradation of performance [49]. It is important to note that the enrolled and tested speaker's language are the same in these experiments. Further, the authors have also pointed out the need for new databases from different languages.

Subsequently, the research work focused on bilingual speakers and performed cross-lingual speaker verification. In the investigation of combining the residual phase cepstral coefficients (RPCC) with Mel-frequency cepstral coefficients (MFCC) work from [237], it is observed that RPCC has improved the performance of traditional speaker verification methods. The residual phase characterizes the glottal closure instants better than the linear prediction models like MFCC. The glottal closure instants are known to contain speaker-specific information [238] [239]. Considering the advantages of residual phase and glottal flow, Wang *et al.* [240] proposed a bilingual speaker identification with RPCC and glottal flow cepstral coefficients (GLFCC) as features. The experiments on NIST SRE 2004

corpus, RPCC features show the highest accuracy when compared to MFCC features.

In [198], Mishra *et al.* examined the language mismatch in speaker verification over i-vector system. When all the parameters are kept consistent, and by changing the language, there is performance degradation in EER by 135%. Also, including a phoneme histogram normalization method using a GMM-UBM system improves the EER by 16%. Li et al. [199] have proposed a deep feature learning for cross-lingual speaker verification in comparison with i-vector based method. Two deep neural networks (DNN) based approaches are proposed with the knowledge of phonemes, which is considered as a linguistic factor. The DNN feature with linguistic factor and PLDA scoring shows better performance than i-vector based method and DNN without linguistic factor.

11.3 X-vector based Speaker Verification system

The X-vector based speaker verification, which is a Deep Neural Network-based approach, proposed by Snyder *et al.* in [33] has the improved performance from data augmentation as suggested in [32]. The model is a feed-forward Deep Neural Network (DNN) which works on cepstral features that are 24-dimensional filter banks and has a frame length of 25 ms with mean-normalization over a sliding window of up to 3 seconds. The model consists of eight layers. The first five layers work on the speech frames, with an added temporal context that is gradually built on through the layers until the last of the five layers. A statistics pooling layer aggregates the outputs and calculates the mean and standard deviation for each input segment. The mean and standard deviation are concatenated and propagated through two segment-level layers and through the last layer, a softmax output layer. The block diagram of x-vector based automatic verification system is show in Figure 11.1

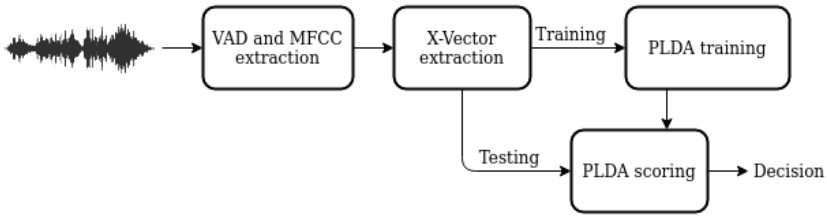


Figure 11.1: Block diagram of X-vector based automatic speaker verification system

The x-vector method is used with two pre-trained variants, one trained on the combined dataset of five Switchboard datasets, SRE datasets from 2004 to 2010, and

the Mixer 6 dataset and the second one is trained on the VoxCeleb 1 and VoxCeleb 2 datasets. The two models are different in multiple directions including the data capturing mechanism, languages spoken in data and variance in acquisition channels. The pre-trained models have been obtained from the Kaldi webpages namely the SRE16 model from http://kaldi-asr.org/models/m3, and the VoxCeleb model from http://kaldi-asr.org/models/m7.

11.3.1 NIST-SRE16 trained model

The NIST-SRE16 pre-trained model uses a total of 15 different datasets, containing a total of 36 different languages. The combined amount of speakers from the Switchboard, SRE, and Mixer datasets totals 91k recordings from over 7k speakers. Data augmentation is done, adding noise and reverberation to the dataset, and combining two augmented copies to the original clean training set. The augmentation of the recording was chosen randomly between four possible types, either augmenting with babble, music, noise, or reverb. Augmenting with babble was done by appending three to seven speakers from the MUSAN speech to the original signal, augmenting with music was selecting a music file randomly from MUSAN, trimmed or repeated to match the duration of the original signal. Noise augmentation was done by adding one-second intervals to the original signal, taken from the MUSAN noises set. Reverb augmentation was done by artificially reverberating via convolution with simulated RIRs.

The SRE16 x-vector model training is employed with with two PLDAs. The first PLDA is trained on the same datasets as the x-vector model trained, but not fitted to the evaluation dataset. As the PLDA is only trained on out-of-domain data, this PLDA is called out-of-domain (OOD) PLDA. The second PLDA (ADT) is fitted to the same datasets and has been adapted to SRE16 data by using the SRE16 major dataset, containing utterances in Cantonese and Tagalog. Therefore, this PLDA is in-domain adapted (ADT) PLDA. The evaluation set of SRE16 is used to test the trained model. The performance of the x-vector method is observed as equal error rate (EER) of 11.73% with OOD PLDA and 8.57% with ADT PLDA.

11.3.2 VoxCeleb trained model

The VoxCeleb model used has been trained on the datasets VoxCeleb 1 and VoxCeleb 2 created by Chung *et al.* in [47] and [50], respectively. The development set of VoxCeleb 1 contains over 140k utterances for 1211 speakers, while the VoxCeleb 2 contains over a million utterances for 6112 speakers. All utterances in VoxCeleb1 are in English but VoxCeleb2 contains multiple languages and have been extracted from videos uploaded to YouTube. The training set size has been increased by using Data Augmentation by adding noise and reverberation to the datasets. In the same fashion as done in Section 11.3.1. The test set of VoxCeleb1

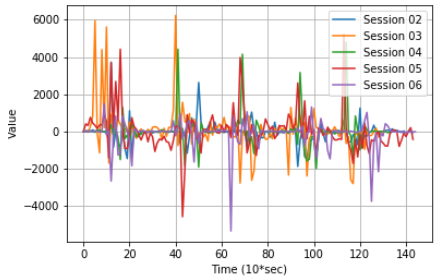


Figure 11.2: A sample signal from SWAN dataset from each session.

with 40 speakers is used to evaluate the training process and the performance is observed as EER of 3.128%.

11.4 Smartphone Multilingual Dataset

The SWAN (Secured access over Wide Area Network) dataset [12] is part of the SWAN project funded by The Research Council of Norway. The data has been gathered using an Apple iPhone6S and has been captured at five different sites. Each site has enlisted 50 subjects in six sessions, where eight individual recordings have been recorded. Depending on the capture site, four of the utterances are in either Norwegian, Hindi, or French, while the remaining four are in English. The utterances spoken are predetermined with alphanumerical speeches. The speakers have pronounced the first utterances in English and then in a national language depending on the site.

The six sessions of data capture are present at each site with a time interval of 1 week to 3 weeks between each session. Session 1 and 2 are captured in a controlled environment with no noise. Session 1 is primarily used to create presentation attack instruments. Therefore, we did not use session 1 data in our experiments. Session 3,4 and 6 are captured in a natural noise environment, and session five is captured in a crowded noise environment. In our experiments, we have enrolled session 2 data in all languages, and other sessions data are used for testing. This way, we can understand the session variance and the impact of noise on ASV methods. A sample of single utterance (sentence 2 in English with duration 14 seconds) is presented in Figure 11.2 indicting the intra-subject variation between different sessions. The Figure 11.2 shows the utterances of the sentence "My account number is *fake account number*" by the same subject in all sessions.

Table 11.1: Results from SRE16-trained X-vector Model with two types of PLDAs and different sessions.

Enrolment language	Test language	S3		S4		S5		S6	
		OOD	ADT	OOD	ADT	OOD	ADT	OOD	ADT
English	English	3.21	3.20	1.65	1.76	4.05	4.15	1.78	1.83
English	Norwegian	6.45	6.65	5.89	5.61	**8.60**	**8.32**	6.16	6.11
English	Hindi	6.83	6.37	5.68	4.96	7.48	7.27	6.33	6.13
English	French	7.76	7.21	5.65	5.08	5.13	4.96	6.13	5.73
Norwegian	Norwegian	3.12	3.21	1.28	1.44	4.98	4.42	1.70	1.77
Norwegian	English	5.56	5.17	3.62	3.42	**8.46**	**7.34**	3.76	2.95
Hindi	Hindi	5.26	4.39	5.01	4.23	4.35	4.46	4.77	4.58
Hindi	English	**7.50**	**7.51**	6.18	5.73	5.45	5.49	5.23	4.72
French	French	5.33	4.32	2.45	2.40	2.62	2.35	1.88	2.06
French	English	6.13	**6.10**	3.41	3.18	**6.44**	5.22	4.63	4.64

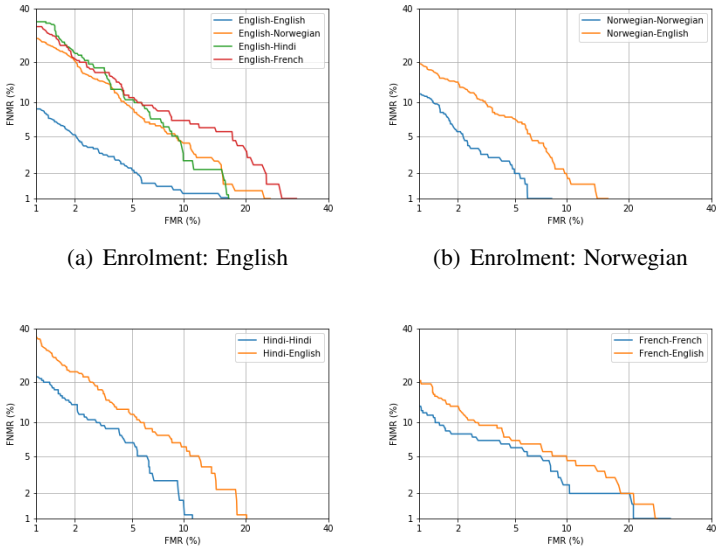
11.5 Experiments and Results

We have four different sets of languages in our dataset, where English is the common language in all the sets. Experiments on four sets of different language combinations are performed. Also, we have five sessions of data capturing in each of the sets. We have followed the same protocol among all the sets by enrolling session two samples and using the rest of the sessions data for testing. To study the cross-lingual speaker recognition results, we have enrolled each language separately and tested the other languages present in that set.

The results are presented using the ISO/IEC standardized metrics for biometric performance [234]. Equal error rate (EER) is the error rate at which the false match rate (FMR) and false non-match rate (FNMR) are equal. We have plotted detection error trade-off (DET) curves, which represent the performance of the recognition of the biometric system in terms of FNMR over FMR.

11.5.1 Experiment 1

The first experiment is carried out on NIST-SRE16 trained model for x-vector extraction and PLDA scoring. This experiment includes two types of PLDA scoring approaches. The first type (OOD PLDA) is an out-of-domain model trained on combined data that contains the Switchboard database, all SREs prior to 2016, and Mixer 6. The second type of PLDA (ADT PLDA) is an in-domain PLDA that is adapted to the SRE16 major partition.



(a) Enrolment: English (b) Enrolment: Norwegian

(c) Enrolment: Hindi (d) Enrolment: French

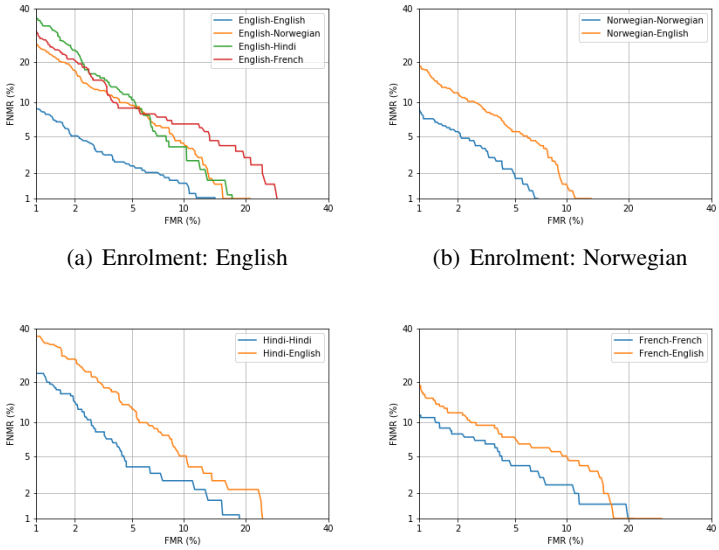
Figure 11.3: DET curves showing the performances of Session 3 with trained model on NIST-SRE16 and out-of-domain adapted PLDA (OOD).

Table 11.1 represents the cross-lingual speaker recognition with English as the enrolment language in all four sessions. The highest error is highlighted among the block of same enrolled language in each PLDA method. It can be clearly seen that the EER values are lower when the enroll language and test languages are the same compared to different languages in test data. Similar results are obtained with Norwegian, Hindi, and French. The highest difference can be observed in the case of English-French combination with a degradation in performance of more than 350% on Session 6 data.

Session 5 has displayed the least accuracy in recognizing speakers among all language combinations. The main reason for this problem could be due to the crowded environment of the data captured. The Figures 11.3 and 11.4 show the plots of DET curves from different languages used in enrolment and testing from Session 3. The error rates can be clearly seen increasing when cross-lingual speaker recognition is performed.

PLDA adaptation

The adaptation of PLDA training does not show a regular trend among different languages. Although the out-of-domain PLDA adaption (OOD) displays higher



(a) Enrolment: English

(b) Enrolment: Norwegian

(c) Enrolment: Hindi

(d) Enrolment: French

Figure 11.4: DET curves showing the performances of Session 3 data and trained on NIST-SRE16 with in-domain adapted PLDA (ADT).

error rates in many cases, in-domain adapted PLDA (ADT) does not improve the performance for some same-language and cross-language evaluations. In the future works, more experiments on different models of OOD and ADT will be studied along with multiple languages included in the data.

11.5.2 Experiment 2

VoxCeleb trained model is used in the second experiment. The PLDA used in this model is trained on VoxCeleb1, and Voxceleb2 combined. A similar protocol from Experiment 1 is followed here also but with only one type of PLDA model. Table 11.2 shows the EER values among different language combination with highest EER value highlighted. The equal error rate is increased in all cases when there is a language mismatch between enrolment and testing. However, it is interesting to observe that the difference in the drop of EER is higher than for Experiment 1.

Figure 11.5 shows the comparison of DET curves between the same language and cross-language speaker recognition from Session 3 of the dataset. It can be clearly seen that the performance of the system has decreased when language mismatch has happened. The difference between the same language and cross-language is much higher in the VoxCeleb model than that of the NIST-SRE16 trained model.

Table 11.2: Results from VoxCeleb X-vector Model from different sessions.

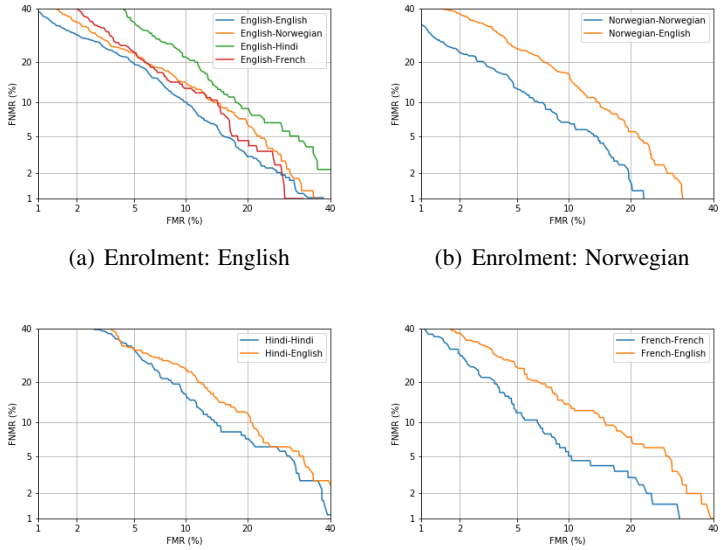
Enrolment language	Test language	S3	S4	S5	S6
English	English	9.90	7.69	10.01	7.99
English	Norwegian	11.83	10.31	**15.01**	10.48
English	Hindi	13.84	13.12	12.75	12.05
English	French	11.21	9.06	11.28	9.46
Norwegian	Norwegian	8.04	6.44	10.91	6.74
Norwegian	English	11.92	9.32	**13.71**	9.55
Hindi	Hindi	12.16	10.68	11.88	10.66
Hindi	English	**14.77**	11.70	13.11	12.72
French	French	7.64	6.58	8.29	6.94
French	English	**11.83**	9.71	8.57	9.41

The speaker recognition accuracy is consistently lower than for the NIST-SRE16 trained model in all the cases. The reason for this could be that the world training dataset in the NIST-SRE16 model contains multiple languages which attributes for cross-lingual speaker recognition robustness. On the other hand, the VoxCeleb2 dataset contains multiple languages, there is a huge variance in data and bias in the number samples per subject which could be reason that limits the ability of the system to recognize different languages in enrolling and testing.

11.6 Conclusion

Behavioral biometric recognition methods have multiple dependencies due to high intra-class variation caused by environmental factors and the human factors impacting the capture process. In the speaker recognition community, dependencies of samples like the text used in the speech and language in which speech is delivered needs to be investigated. The dependency due to language has been a problem when there is a mismatch between enrolment and tested language. In this work, we have focused on evaluating the problem of language mismatch on the state-of-the-art speaker recognition method, namely the x-vector method, which uses a deep neural network-based approach. We have chosen a multilingual dataset with four different languages and four different sessions. For the world training dataset, we included two popular publicly available datasets NIST-SRE16 and VoxCeleb.

The experiments on cross-lingual speaker recognition displayed the performance degradation when there is a mismatch in languages in enrolment and testing. Fur-



(a) Enrolment: English

(b) Enrolment: Norwegian

(c) Enrolment: Hindi

(d) Enrolment: French

Figure 11.5: DET curves showing the performances of Session 3 data and trained on VoxCeleb data.

ther, the dependency on the languages included in the world training dataset is observed. If there are multiple languages used in the world training dataset, which is the case of NIST-SRE16, performance degradation is less compared to the one language model VoxCeleb. In future works, a speaker recognition approach is implemented to overcome the problem of language dependency.

Chapter 12

Article 7: Smartphone audio replay attacks dataset

Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. Smartphone audio replay attacks dataset. In *2021 9th IEEE International Workshop on Biometrics and Forensics(IWBF)*. IEEE, 2021

12.1 Abstract

Smartphone based biometric applications are increasing exponentially in recent years. The challenges due to presentation attacks in biometrics have emerged to cause a potential vulnerability, limiting the reliability of biometrics for secure applications. In speaker recognition, audio replay attacks have demonstrated a severe threat to automatic speaker verification (ASV) systems. Alongside, the difference in language for enrolment and testing has displayed some impact on speaker recognition. In this direction, we have created a novel audio replay attack dataset for four different languages using smartphones as playback and recording devices. We have collected data in two different scenarios where the attack recording sensor and bona fide sensor are the same and different. The captured dataset is used for testing the vulnerability on both state-of-the-art speaker recognition method and commercial-off-the-shelf (COTS) method from VeriSpeak. The baseline presentation attack detection methods are benchmarked on replay attacks in a cross-smartphone scenario. The results show that the replay attacks indicate a severe threat towards the ASV methods, especially in the cross-smartphone scenario.

12.2 Introduction

Smartphone applications have been growing enormously across different areas of data processing. The usage of mobile data processing includes sensitive information related to the user, and this requires security by authorization. Therefore, biometric identification has come to play in most recent smartphones. Face, fingerprint, and iris biometric recognition are prominently used to provide secure access to smartphones. Also, banking applications use the device in-built biometric recognition for smooth banking transactions. However, the presentation attacks cause a severe problem to the optimal performance of embedded biometric systems.

Presentation attacks are defined as the presentation of artefacts to a biometric capture device to interfere with the biometric recognition. In speaker recognition, well-known presentation attacks are replay attacks, voice impersonation, voice conversion, and speech synthesis. Among these, replay attacks are performed by playing back an audio sample to the biometric capture device. A digital copy of the speech sample is required for carrying out this attack. In a general scenario, a speaker device is enough to perform this attack. The biometric methods embedded into smartphone devices use the default microphone to record the speech data. However, the speakers present in these devices can be used as playback devices to play a speech sample to another biometric device.

In this work, we examine the situation where a smartphone based biometric system is attacked using other smartphone speakers. We have used smartphone biometric data and developed a set of audio replay attacks using smartphones as both playback and recording devices. The rest of the paper is described as follows: Section 12.3 provides a brief overview of the related works about the replay attack datasets. Section 12.4.2 describes the smartphone audio replay attack dataset developed in this work. In Section 12.5, we present the ASV methods and baseline attack detection methods that are used to evaluate the vulnerability of speaker recognition systems with regards to replay attacks. Section 12.6 discuss the experiments and results on the developed replay attack dataset. Section 12.7 concludes this work.

12.3 Literature Review

Audio replay attacks are a trending way of interfering with the performance of a speaker recognition system. With the availability of a digital copy of the target speaker's speech, an audio sample can be played back in front of a biometric system to perform a replay attack. Over the years, there are multiple approaches to examine the impact of replay attacks and proposed countermeasures. In this section, we discuss some of those approaches and show promising results.

Janicki *et al.* examined the vulnerability on ASV methods caused by replay attack in comparison to other attacks like voice conversion and speech synthesis [241]. The replay attacks captured in three different conditions are evaluated over six ASV methods, and the results show an increase in EER, which is higher in the case of replay attacks compared to other attacks. Countermeasures using far-field FFD and local binary patterns were proposed, and it is observed that replay attacks captured in the anechoic environment are challenging to detect.

The Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof) has provided an extensive dataset for research and development of countermeasures for replay attacks. The 2017 ASVspoof challenge [232] released a replay attack dataset with various attack configurations. The baseline ASV method using Mel-frequency cepstral coefficient (MFCC) features and a 512-component UBM model display vulnerability due to replay attacks. The dataset also provided two baseline countermeasures based on CQCC features trained on GMM models. The best performing submission to this challenge detected the attacks with an EER of 6.73%. The authors have also published metadata analysis on ASVspoof 2017 replay attacks corpus version 2.0 [242]. A detailed discussion on the different replay configurations with their impact on ASV systems in this work and from the results, the authors have concluded that the high-quality speakers and microphone setups in a studio or anechoic rooms cause a major threat to the performance of ASV methods.

A survey on the replay attack detection methods is discussed in [243]. This work has evaluated the ASVspoof 2017 replay attacks on different countermeasures and discussed the limitations and challenges. In the next series of ASVspoof challenges, replay attacks are evaluated separately, coining as physical access (PA) attacks [40]. The baseline systems used in this work are LFCC and CQCC features with a standard GMM back-end classifier. The best performing submission to this challenge gives an output of EER = 0.39% on the PA scenario.

This work considers the replay attack configurations under three different scenarios:

1. Using widely available smartphones as both playback and attack devices.
2. Evaluating the impact of replay attacks on the language of the speech data.
3. Observing the impact of replay attack scenario when recording device in bona fide and presentation attack is the same and different.

By pointing out these questions, this work has made the following contributions.

1. A novel smartphone audio replay attack dataset with multiple latest smartphones in four different languages.
2. The evaluation of smartphone biometrics' vulnerability based on speaker recognition towards the replay attacks performed using different smartphones.
3. Benchmarked the audio replay attacks on the baseline presentation attack detection methods used in the ASVspoof attack detection challenge.
4. Evaluation and comparison of attack impact and detection in two cases of the same sensor in bona fide and PA capture.

12.4 Smartphone Replay attacks dataset

This section discusses the replay attack data capturing setup, playback-record configurations, and the bona fide dataset.

12.4.1 Data Capturing Setup

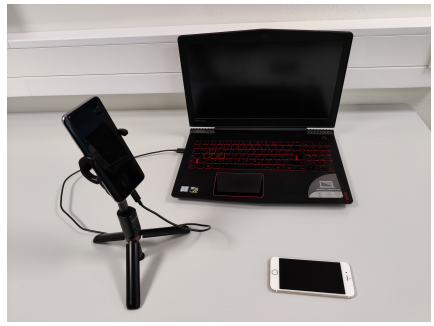


Figure 12.1: Audio replay attack setup.

We have used six different smartphones in the process of creating audio replay attacks. They include four Apple devices (iPhone 6S, iPad Pro, iPhone 10, iPhone 12) and two Samsung devices (S8 and S10). The data capturing setup is presented in Figure 12.1 where a tripod holds a recording device in front of a playback device. We have collected the replay data in an office room with no background noise. A laptop is used to activate the microphone and speaker of the smartphones via USB cables or Bluetooth. Thus, instructions are given to a smartphone from a laptop to play the audio sample on one smartphone and simultaneously record it on another smartphone. This way, the recorded audio samples are directly stored in the laptop.

The playback-record configurations are designed depending on the bona fide data samples. The recording device is kept the same for the initial five configurations

as iPhone 6S. The reason behind this is that the bona fide dataset is also captured with the same device. Therefore, this will allow us to evaluate the performance of different playback devices. In the next five configurations, the rest of the devices are chosen as presented in Table 12.1. The dataset is made publicly available for research purposes [1].

Table 12.1: Replay attack setups

ID	Playback device	Recording device
RP1	Samsung S8	iPhone 6S
RP2	Samsung S10	iPhone 6S
RP3	iPad Pro	iPhone 6S
RP4	iPhone 10	iPhone 6S
RP5	iPhone 12	iPhone 6S
RP6	iPhone 10	iPhone 12
RP7	Samsung S10	iPhone 10
RP8	Samsung S10	iPad Pro
RP9	Samsung S8	iPhone 10
RP10	Samsung S8	iPad Pro

12.4.2 SWAN dataset

In our experiments, we used the SWAN (Secured access over Wide Area Network) dataset [12]. The dataset contains smartphone audio-visual biometric data of 50 subjects captured using an iPhone6S at five different locations. There are six sessions of data from each location. However, we have taken only audio data from session 1, which is extracted from audio-visual samples. Each subject provided eight recordings, among which the first four are in English, and the rest are in either Norwegian, Hindi, or French, depending on the location. The utterances are predetermined and consist of alphanumerical speeches. We have used English audio data from one location and each other language data from corresponding locations for convenience. Therefore, we have data on four different languages with 50 subjects each and each subject speaking four utterances.

SWAN dataset also contains replay presentation attack data using Logitech high-quality loudspeaker and recording using iPhone 6S (protocol PA.V.4 in [12]). This data is used as training data for the PAD methods in our work.

[1] Access to dataset: https://forms.gle/EV5ur4jbw52RgM3v8

12.4.3 Replay Attack Data

The speech samples are played individually on the playback device and recorded from the microphone on the recording device continuously. To cope with the latency of the devices, two seconds pause is provided between the capture of each sample. A similar procedure is followed using all the configurations mentioned in Table 12.1. The spectrograms of a bona fide sample and corresponding replay attack sample are presented in Fig. 12.2.

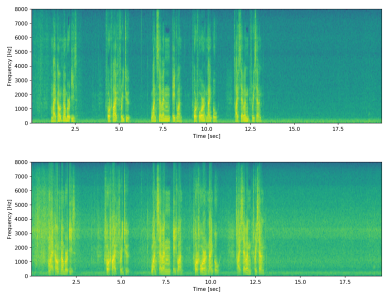


Figure 12.2: Spectrograms of the data samples. Top: Bona fide. Bottom: Replay attack.

12.5 Baseline Methods

The replay attacks are evaluated for the vulnerability towards the Automatic Speaker Verification (ASV) systems and tested on Presentation Attack Detection (PAD) methods. In this section, we present the baseline ASV systems and PAD methods used in our experiments.

12.5.1 Automatic Speaker Verification Methods

We have chosen two ASV methods, including a state-of-the-art method and a Commercial-Off-The-Shelf (COTS) method from VeriSpeak. A brief description of these methods is presented below.

X-vector based Speaker Verification system

The X-vector speaker verification is a feed-forward Deep Neural Network (DNN) approach proposed by Snyder *et al.* in [33]. This approach has proven to improve the ASV system's performance using data augmentation as suggested in [32]. The model uses cepstral features with 24-dimensional filter banks and has a frame length of 25 ms with mean-normalization over a sliding window of up to 3 seconds. The neural network model consists of eight layers, with the first five layers work on the speech frames and an added temporal context that is gradually built on through the layers until the last of the five layers. A statistics pooling layer

aggregates the outputs and calculates the mean and standard deviation for each input segment. The mean and standard deviation are concatenated and propagated through two segment-level layers and the last layer, a softmax output layer.

The X-vectors of the enrol and test samples are compared using the Probabilistic Linear Discriminant Analysis approach proposed in [133]. For the X-vector extractor and PLDA training models, the Kaldi automatic speech recognition toolkit[2] pre-trained models are used which are trained on the VoxCeleb dataset [233].

VeriSpeak method

To examine an operationally deployed biometric system's potential vulnerability towards an attack, evaluating an up-to-date commercial method is required. Therefore, in our work, a standalone speaker recognition approach called VeriSpeak[2] is used. VeriSpeak speaker verification technology is designed for biometric system developers and integrators. The speaker recognition algorithm assures system security by checking voice authenticity. The speaker modelling and recognition methodology used in this method are unknown. The inbuilt liveness detection or presentation attack detection algorithm used in this method is not specified in the release document. The text-independent feature extraction is activated while using the VeriSpeak SDK.

12.5.2 Presentation Attack Detection Methods

The baseline PAD methods provided in the ASVSpoof 2019 challenge are used in our experiments [40]. There are two PAD methods provided by ASVSpoof 2019 evaluation kit. These methods are based on two cepstral coefficients in the front-end, namely Linear Frequency Cepstral Coefficients (LFCC) and Constant Q Cepstral Coefficients (CQCC). The back-end comprises Gaussian Mixture Models (GMM).

The LFCC features are similar to the popular Mel-frequency cepstral coefficients (MFCCs), but the filters are placed linearly in equal sizes. The LFCC features were proposed for synthetic-detection in [194]. LFCC features are extracted on speech signals with a frame length of 25ms and a 20-channel linear filter bank, which includes 19 cepstral coefficients, a zeroth coefficient, the static, delta, and delta-delta coefficients.

The CQCC features are created with a maximum frequency of fs/2, where fs = 16kHz is the sampling frequency. The minimum frequency is set to $fs/2/2^9$ $15Hz$ (9 being the number of octaves). [195]. The number of bins per octave is set to 96.

[2]Kaldi ASR toolkit: http://kaldi-asr.org/
[2]VeriSpeak: https://www.neurotechnology.com/verispeak.html

Re-sampling is applied with a sampling period of 16. CQCC features dimension is set to 29 coefficients + 0th, with the static, delta, and delta-delta coefficients.

In the back-end, 2-class GMMs are trained on the bona fide and attack speech utterances of the training dataset, respectively. We use 512-component models trained with an expectation-maximization (EM) algorithm with random initialization. The attack detection score is computed as the log-likelihood ratio for the test utterance given bona fide and the replay attacks speech models.

12.6 Experiments and Results

12.6.1 Evaluation Metrics

The performance of speaker recognition systems and presentation attack detection (PAD) methods are evaluated using ISO/IEC standard biometric metrics [244].

- False Match Rate (FMR) is the proportion of the completed biometric non-mated comparison trials that result in a false match, and False Non-Match Rate (FNMR) is the proportion of the completed biometric mated comparison trials that result in a false non-match.
- Impostor-Attack Match Rate (IAMPR) is the proportion of impostor attack samples (replay attacks) that are matched with bona fide samples. To compare ASV methods' performance, we have fixed FMR at 0.1% and presented FNMR and IAPMR for zero-effort impostors and attacks, respectively.
- Attack Presentation Classification Error Rate (APCER) is the proportion of attack presentations that are incorrectly classified as bona fide presentations, and Bona fide Presentation Classification Error Rate (BPCER) is the ratio of bona fide presentation incorrectly classified as attacks. In this work, we have presented the BPCER of PAD methods by fixing APCER at 10%.

In addition to ISO/IEC metrics mentioned above, the Detection Equal Error Rate (D-EER) is presented for PAD methods which is the attack detection rate at the working point, when APCER and BPCER are equal.

12.6.2 Vulnerability analysis

The impact of replay attacks on the ASV systems is examined by computing verification scores of bona fide samples against replay attack samples. Protocols are created for calculating verification scores of bona fide samples against zero-effort impostors and replay attacks. In the case of zero-effort impostors, each speech sample of a subject is paired with each of the other samples from the same subject

(mated scores), and samples from other subjects (non-mated scores) are chosen randomly. This results in bona fide scores and zero-impostors scores, respectively. The replay attacks are paired with corresponding bona fide audio samples, and verification scores are obtained. This protocol gives the attack scores.

The verification protocols are created for each of the replay attack configurations and four different languages separately. To avoid the dependency of the language, we enrolled and tested the biometric samples in sample language. Table 12.2 presents the ASV systems performance with no attacks i.e. Zero-Effort impostors. The FMR is fixed at 0.1%, and FNMR values for each language and two ASV methods are computed. For the X-vector method, the threshold for FMR at 0.1% is computed on individual languages, where the VeriSpeaker method's threshold is obtained from the documentation provided by NeuroTechnology (which is 36 for FAR=0.1%).

Table 12.2: FNMR% at FMR = 0.1% for Zero-effort impostors

ASV method	Language			
	English	Norwegian	French	Hindi
X-vector	9	15.33	21	11.48
VeriSpeak	36.66	22.66	25	7.78

The speaker verification results with zero-effort impostors show various levels of performance for different languages and ASV methods. It is necessary to observe that the state-of-the-art approach performed better than the commercial methods. Also, the commercial VeriSpeak method displayed the least performance in the English language, where FNMR is 36.66%. The possible reason for this result is that the audio sample quality in the SWAN dataset. The VeriSpeak ASV method is set up with default settings irrespective of data quality. We assume that this caused the drop in speaker verification performance of data in the English language.

Tables 12.3 and 12.4 presents the IAPMR values with fixed FMR threshold similar to zero-effort impostors. IAPMR determines the number of attack samples being matched with the target bona fide samples. It is observed from results that replay attacks show high vulnerability in both ASV methods. It is also interesting to notice that languages have a minimum impact on the replay attacks. The first five attack setups have the same recording device as the bona fide recording device, the iPhone 6S. The last five have different recording devices but display similar vulnerabilities due to attacks.

The X-vector ASV method allows most of the attack samples to match with bona fide samples except the setup of RP10: playback device is Samsung S8 and recording device is Apple iPad pro. It is noted that the sound of speech data in the RP10

Table 12.3: IAPMR% at FMR = 0.1% for X-vector method

Attack setup	Language			
	English	Norwegian	French	Hindi
RP1	100	99.51	99.50	97.28
RP2	99.50	100	98.53	97.83
RP3	100	98.55	99	97.81
RP4	99.51	99.03	99.50	99.45
RP5	100	99.01	99.50	98.89
RP6	99.02	100	98.54	98.91
RP7	100	98.07	99.01	98.37
RP8	99.50	100	99	99.44
RP9	99.50	100	99.50	98.89
RP10	25.62	25.75	26.06	25.53

setup is comparatively low due to some technical problem. This resulted in the X-vector system discarding many samples as speech data with no voice. Therefore, only 25% of the replay attack data match with bona fide data in this configuration.

The VeriSpeak method displays slightly lower IAPMR rates when the same recording device is used to capture attacks and bona fide samples (RP1 to RP5). This may prove that the commercial ASV method is relatively better than the state-of-the-art approach. However, the overall performance of the COTS method shows that replay attacks cause high vulnerability.

Table 12.4: IAPMR% at FMR = 0.1% for VeriSpeak method

Attack setup	Language			
	English	Norwegian	French	Hindi
RP1	99.50	98.51	99.50	98.36
RP2	99	99.51	98.53	98.37
RP3	95.01	94.17	98.52	98.91
RP4	94.31	95.38	98.33	98
RP5	99.50	98.02	98.51	100
RP6	95.09	95.56	95.14	98.03
RP7	99.50	97.59	99.01	98.38
RP8	99.54	99.50	99.50	99.44
RP9	99.52	98.52	99.51	99.45
RP10	99	99	98.53	98.91

Figure 12.3 shows the distribution of bona fide, zero-effort impostors and replay

attack scores for the VeriSpeak method. The threshold at which the FMR = 0.1% is also plotted, which displays the overlap of attack scores with bona fide scores. In the X-vector methods, the threshold for FMR = 1.0% is obtained for each of the attack setup and language. Therefore, we have not included the distributions of various setups for the X-vector method in this paper.

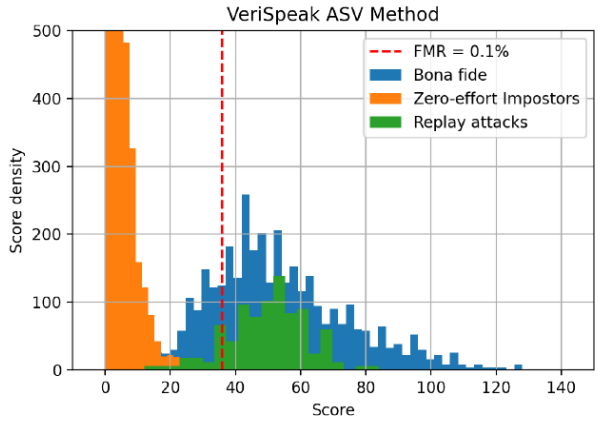


Figure 12.3: Vulnerability Evaluation of VeriSpeak method.

12.6.3 Replay attack detection

The replay attack detection is performed using two baseline methods provided in ASVSpoof 2019 evaluation plan. The GMM models are trained on the bona fide and replay attack PA.V.4 data of the first 30 subjects from the SWAN dataset. For testing the replay attacks captured in this work, the speech samples from the last 20 subjects are evaluated. The loglikelihood scores are calculated for each of the bona fide and attack data for computing D-EER and BPCER at APCER = 10%.

Table 12.5 presents the results from the PAD methods on all the attack setup. The attack setups from RP1 to RP5 have the same recording device as that of the attack data used in training. The CQCC method detected replay attacks better than the LFCC method. However, in the case of RP4 and RP5, CQCC is not able to detect replay attacks. This behaviour is not unknown, and further conclusions can be deduced with a detailed evaluation of the CQCC method of replay attacks. From RP6 to RP10, the recording device in bona fide data and attack data are different. Therefore, it is observed from the results that the same recording device in training and testing displays better PAD performance.

It is also noticed that when the recording device and playback device are from the

Table 12.5: D-EER% and BPCER_10% (BPCER @ APCER = 10%) for baseline PAD methods

Attack setup	LFCC		CQCC	
	D-EER%	BPCER_10%	D-EER%	BPCER_10%
RP1	39.92	49.05	10.23	10.23
RP2	35.42	86.93	15.34	20.45
RP3	0	0	0	0
RP4	5.25	2.38	48	68.45
RP5	4.55	1.71	50	90.15
RP6	0	0	0	0
RP7	38.06	79.17	31.06	56.63
RP8	14.58	21.59	23.10	36.93
RP9	43.56	73.48	43.18	78.41
RP10	38.82	67.99	41.29	78.98

same manufacturer (in the case of RP3, RP4, RP5, and RP6), the PAD methods performed better. However, when the device manufacturer is not the same, the performance is not consistent. For example, the RP8 and RP10 configurations have the same recording device (iPad Pro) but different playback devices from the same manufacturer. The attack detected methods performed better in RP8 than in RP10. This observation also concludes that change in playback device can impact the PAD methods.

12.7 Conclusion

Smartphone biometrics have emerged into daily usage of biometric-based authentication for multiple purposes. The presentation attacks pose a major problem in the proper functioning of biometric algorithms embedded in smartphones. Among these, replay attacks are easy to perform and displayed high false match rates. In this work, we have created a smartphone-based replay attack dataset in four different languages. Six different smartphones are used with two different scenarios of the same and different recording devices to bona fide data. Two ASV methods are evaluated for the impact of replay attacks, and results show high vulnerability caused by the replay attacks. Further, we have performed presentation attack detection using baseline countermeasures, and it is noted that the attacks are relatively easy to detect when the recording device of replay attacks is the same as that of bona fide data.

Acknowledgment

This work is carried out under the SWAN (Secured access over Wide Area Network) project funded by the Research Council of Norway (Grant No. IKTPLUSS 248030/O70)

Bibliography

[1] Petar S Aleksic and Aggelos K Katsaggelos. An audio-visual person identification and verification system using faps as visual features. *Works. Multimedia User Authentication, Santa Barbara, CA*, 2003.

[2] Girija Chetty and Michael Wagner. Multi-level liveness verification for face-voice biometric authentication. In *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, pages 1–6. IEEE, 2006.

[3] Daksha Yadav, Naman Kohli, James S Doyle, Richa Singh, Mayank Vatsa, and Kevin W Bowyer. Unraveling the effect of textured contact lenses on iris recognition. *IEEE Transactions on Information Forensics and Security*, 9(5):851–862, 2014.

[4] Xiaofu He, Shujuan An, and Pengfei Shi. Statistical texture analysis-based approach for fake iris detection using support vector machines. In *International Conference on Biometrics*, pages 540–546. Springer, 2007.

[5] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 30107-1. Information Technology - Biometric presentation attack detection - Part 1: Framework*. International Organization for Standardization, 2016.

[6] Conrad Sanderson and Kuldip K Paliwal. Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449–480, 2004.

[7] Tsuhan Chen. Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18(1):9–21, 2001.

[8] Enrique Bailly-Bailliére, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad

Popovici, Fabienne Porée, Belen Ruiz, and Jean-Philippe Thiran. The banca database and evaluation protocol. In *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication*, AVBPA'03, page 625–638, Berlin, Heidelberg, 2003. Springer-Verlag.

[9] Niall A. Fox, Brian A. O'Mullane, and Richard B. Reilly. Valid: A new practical audio-visual database, and comparative results. In Takeo Kanade, Anil Jain, and Nalini K. Ratha, editors, *Audio- and Video-Based Biometric Person Authentication*, pages 777–786, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[10] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[11] Javier Ortega-Garcia, Julian Fierrez, Fernando Alonso-Fernandez, Javier Galbally, Manuel R Freire, Joaquin Gonzalez-Rodriguez, Carmen Garcia-Mateo, Jose-Luis Alba-Castro, Elisardo Gonzalez-Agulla, Enrique Otero-Muras, et al. The multiscenario multienvironment biosecure multimodal database (bmdb). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1097–1111, 2009.

[12] Raghavendra Ramachandra, Martin Stokkenes, Amir Mohammadi, Sushma Venkatesh, Kiran Raja, Pankaj Wasnik, Eric Poiret, Sébastien Marcel, and Christoph Busch. Smartphone multi-modal biometric authentication: Database and evaluation. *arXiv preprint arXiv:1912.02487*, 2019.

[13] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007.

[14] A. Ross and A. K. Jain. Multimodal biometrics: An overview. In *2004 12th European Signal Processing Conference*, pages 1221–1224, 2004.

[15] Sébastien Marcel, Mark S Nixon, Julian Fierrez, and Nicholas Evans. *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019.

[16] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC TR 24741 Biometrics – Overview and application*. International Organization for Standardization, 2020.

[17] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004.

[18] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[19] Al-Amin Bhuiyan and Chang Hong Liu. On face recognition using gabor filters. *World academy of science, engineering and technology*, 28, 2007.

[20] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[21] Hareesh Mandalapu, Aravinda Reddy P N, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, S. R. Mahadeva Prasanna, and Christoph Busch. Audio-visual biometric recognition and presentation attack detection: A comprehensive survey. *IEEE Access*, 9:37431–37455, 2021.

[22] Maycel-Isaac Faraj and Josef Bigun. Audio–visual person authentication using lip-motion from orientation maps. *Pattern recognition letters*, 28(11):1368–1382, 2007.

[23] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[25] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[26] John Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009.

[27] Inmaculada Tomeo-Reyes and Vinod Chandran. Iris based identity verification robust to sample presentation security attacks. *International Journal of Information Science and Intelligent System*, 2(1):27–41, 2013.

[28] Adam Czajka and Kevin W Bowyer. Presentation attack detection for iris recognition: An assessment of the state of the art. *arXiv preprint arXiv:1804.00194*, 2018.

[29] Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech communication*, 54(4):543–565, 2012.

[30] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[31] Simon Prince, Peng Li, Yun Fu, Umar Mohammed, and James Elder. Probabilistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157, 2012.

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

[33] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017-August, pages 999–1003, 2017. Cited By :202.

[34] Robert W Frischholz and Ulrich Dieckmann. Biold: a multimodal biometric identification system. *Computer*, 33(2):64–68, 2000.

[35] Seyed Omid Sadjadi, Craig S Greenberg, Elliot Singer, Douglas A Reynolds, Lisa Mason, and Jaime Hernandez-Cordero. The 2019 nist audio-visual speaker recognition evaluation. *Proc. Speaker Odyssey (submitted), Tokyo, Japan*, 2020.

[36] EasyPASS-Grenzkontrolle einfach und schnell. http://www.bundespolizei.de/DE/01Buergerservice/Automatisierte-Grenzkontrolle/EasyPass/_easyPass_anmod.html., 2014.

[37] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC FDIS 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting*. International Organization for Standardization, 2017.

[38] Raghavendra Ramachandra and Christoph Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 50(1):1–37, 2017.

[39] Hareesh Mandalapu, Aravinda Reddy P N, Raghavendra Ramachandra, K Sreenivasa Rao, Pabitra Mitra, S R Mahadeva Prasanna, and Christoph Busch. Multilingual audio-visual smartphone dataset and evaluation, 2021.

[40] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[41] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[42] Ramachandra Raghavendra and Christoph Busch. Robust scheme for iris presentation attack detection using multiscale binarized statistical image features. *IEEE Transactions on Information Forensics and Security*, 10(4):703–715, 2015.

[43] Deng Cai, Xiaofei He, and Jiawei Han. Efficient kernel discriminant analysis via spectral regression. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 427–432. IEEE, 2007.

[44] Ana Carolina Lorena and André CPLF De Carvalho. Building binary-tree-based multiclass classifiers using separability measures. *Neurocomputing*, 73(16-18):2837–2845, 2010.

[45] David Yambay, Benedict Becker, Naman Kohli, Daksha Yadav, Adam Czajka, Kevin W Bowyer, Stephanie Schuckers, Richa Singh, Mayank Vatsa, Afzel Noore, et al. Livdet iris 2017—iris liveness detection competition 2017. In *Biometrics (IJCB), 2017 IEEE International Joint Conference on*, pages 733–741. IEEE, 2017.

[46] Aythami Morales, Julian Fierrez, Javier Galbally, and Marta Gomez-Barrero. Introduction to iris presentation attack detection. In *Handbook of Biometric Anti-Spoofing*, pages 135–150. Springer, 2019.

[47] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[48] Xugang Lu and Jianwu Dang. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech communication*, 50(4):312–322, 2008.

[49] Roland Auckenthaler, Michael J Carey, and John SD Mason. Language dependency in text-independent speaker verification. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 441–444. IEEE, 2001.

[50] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[51] Petar S Aleksic and Aggelos K Katsaggelos. Audio-visual biometrics. *Proceedings of the IEEE*, 94(11):2025–2044, 2006.

[52] Kai Li. Identity authentication based on audio visual biometrics: A survey. *URL: http://www. eecs. ucf. edu/˜ kaili/pdfs/survey_avbiometr ics. pdf*, 2013.

[53] Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, 2010.

[54] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[55] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernockỳ, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *2012 IEEE International Conference on Multimedia and Expo Workshops*, pages 635–640. IEEE, 2012.

[56] Stéphane Pigeon and Luc Vandendorpe. The m2vts multimodal face database (release 1.00). In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 403–409. Springer, 1997.

[57] Conrad Sanderson and Brian C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In Massimo Tistarelli and Mark S. Nixon, editors, *Advances in Biometrics*, pages 199–208, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[58] Denis Burnham, Dominique Estival, Steven Fazio, Jette Viethen, Felicity Cox, Robert Dale, Steve Cassidy, Julien Epps, Roberto Togneri, Michael Wagner, et al. Building an audio-visual corpus of australian english: large corpus collection with an economical portable and replicable black box. ISCA, 2011.

[59] Denis Burnham, Eliathamby Ambikairajah, Joanne Arciuli, Mohammed Bennamoun, Catherine T Best, Steven Bird, Andrew R Butcher, Steve Cassidy, Girija Chetty, Felicity M Cox, et al. A blueprint for a comprehensive australian english auditory-visual speech corpus. In *HCSNet Workshop on Designing the Australian National Corpus*, pages 96–107. Cascadilla Proceedings Project, 2009.

[60] Michael Wagner, Dat Tran, Roberto Togneri, Phil Rose, David Powers, Mark Onslow, Debbie Loakes, Trent Lewis, Takaaki Kuratate, Yuko Kinoshita, et al. The big australian speech corpus (the big asc). In *SST 2010, Thirteenth Australasian International Conference on Speech Science and Technology*, pages 166–170. ASSTA, 2011.

[61] Thomas Bräunl, Brendan McCane, Mariano Rivera, and Xinguo Yu. *Image and Video Technology: 7th Pacific-Rim Symposium, PSIVT 2015, Auckland, New Zealand, November 25-27, 2015, Revised Selected Papers*, volume 9431. Springer, 2016.

[62] André Anjos, Laurent El-Shafey, Roy Wallace, Manuel Günther, Christopher McCool, and Sébastien Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1449–1452, 2012.

[63] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[64] Mikhail I Gofman, Sinjini Mitra, Tsu-Hsiang Kevin Cheng, and Nicholas T Smith. Multimodal biometrics for enhanced mobile device security. *Communications of the ACM*, 59(4):58–65, 2016.

[65] SPEECHPRO. https://speechpro-usa.com/product/voice_authentication/voicekey-onepass, 2019.

[66] Elie Khoury, Manuel Günther, Laurent El Shafey, and Sébastien Marcel. On the improvements of uni-modal and bi-modal fusions of speaker and face recognition for mobile biometrics. Technical report, Idiap, 2013.

[67] RESPECT. http://www.respect-project.eu/team.html, 2019.

[68] Claude C Chibelushi, F Deravi, and JS Mason. Voice and facial image integration for person recognition. 1994.

[69] Souheil Ben-Yacoub, Yousri Abdeljaoued, and Eddy Mayoraz. Fusion of face and speech data for person identity verification. *IEEE transactions on neural networks*, 10(5):1065–1074, 1999.

[70] Timothy J Hazen, Eugene Weinstein, Ryan Kabir, Alex Park, and Bernd Heisele. Multi-modal face and speaker identification on a handheld device. In *proceedings of the Workshop on Multimodal User Authentication*, pages 113–120. Citeseer, 2003.

[71] Pierre Jourlin, Juergen Luettin, Dominique Genoud, and Hubert Wassner. Integrating acoustic and labial information for speaker identification and verification. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[72] Timothy Wark, Sridha Sridharan, and Vinod Chandran. Robust speaker verification via fusion of speech and lip modalities. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 6, pages 3061–3064. IEEE, 1999.

[73] Timothy Wark, Sridha Sridharan, and Vinod Chandran. The use of temporal speech and lip information for multi-modal speaker identification via multi-stream hmms. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 4, pages 2389–2392. IEEE, 2000.

[74] Upendra V Chaudhari, Ganesh N Ramaswamy, Gerasimos Potamianos, and Chalapathy Neti. Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–9. IEEE, 2003.

[75] Ulrich Dieckmann, Peter Plankensteiner, Ralf Schamburger, Bernhard Fröba, and Sebastian Meller. Sesam: A biometric person identification system using sensor fusion. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 301–310. Springer, 1997.

[76] Girija Chetty and Michael Wagner. Liveness verification in audio-video speaker authentication. In *in Proc. 10th ASSTA conference*. Citeseer, 2004.

[77] Enrique Argones Rúa, Hervé Bredin, Carmen García Mateo, Gérard Chollet, and Daniel González Jiménez. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. *Pattern Analysis and Applications*, 12(3):271–284, 2009.

[78] Elhocine Boutellaa, Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Audiovisual synchrony assessment for replay attack detection in talking face biometrics. *Multimedia Tools and Applications*, 75(9):5329–5343, 2016.

[79] Walid Karam, Hervé Bredin, Hanna Greige, Gérard Chollet, and Chafic Mokbel. Talking-face identity verification, audiovisual forgery, and robustness issues. *EURASIP Journal on Advances in Signal Processing*, 2009(1):746481, 2009.

[80] Farzin Deravi. Audio-visual person recognition for security and access control. 1999.

[81] Christian Micheloni, Sergio Canazza, and Gian Luca Foresti. Audio–video biometric recognition for non-collaborative access granting. *Journal of Visual Languages & Computing*, 20(6):353–367, 2009.

[82] Niall Fox and Richard B Reilly. Audio-visual speaker identification based on the use of dynamic audio and visual features. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 743–751. Springer, 2003.

[83] Chenxi Yu and Lin Huang. Biometric recognition by using audio and visual feature fusion. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 173–178. IEEE, 2012.

[84] C. C. Chibelushi, F. Deravi, and J. S. Mason. Audio-visual person recognition: an evaluation of data fusion strategies. In *European Conference on Security and Detection, 1997. ECOS 97.*, pages 26–30, 1997.

[85] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-4:2008. Information Technology – Biometric Performance Testing and Reporting – Part 4: Testing methodologies for technology and scenario evaluation*. International Organization for Standardization and International Electrotechnical Committee, 2008.

[86] Lawrence Rabiner. Fundamentals of speech recognition. *Fundamentals of speech recognition*, 1993.

[87] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 29794-1:2009 Information Technology - Biometric Sample Quality - Part 1: Framework*. International Organization for Standardization, 2009.

[88] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland. Multimodal person recognition using unconstrained audio and video. In *Proceedings, International Conference on Audio-and Video-Based Person Authentication*, pages 176–181. Citeseer, 1999.

[89] Ara V Nefian, Lu Hong Liang, Tieyan Fu, and Xiao Xing Liu. A bayesian approach to audio-visual speaker identification. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 761–769. Springer, 2003.

[90] Dhaval Shah, Kyu J Han, and Shrikanth S Narayanan. A low-complexity dynamic face-voice feature fusion approach to multimodal person recognition. In *2009 11th IEEE International Symposium on Multimedia*, pages 24–31. IEEE, 2009.

[91] Samy Bengio. Multimodal authentication using asynchronous hmms. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 770–777. Springer, 2003.

[92] Richard M Jiang, Abdul H Sadka, and Danny Crookes. Multimodal biometric human recognition for perceptual human–computer interaction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):676–681, 2010.

[93] Xuran Zhao, Nicholas Evans, and Jean-Luc Dugelay. Multi-view semi-supervised discriminant analysis: A new approach to audio-visual person recognition. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 31–35. IEEE, 2012.

[94] Elie Khoury, Laurent El Shafey, Christopher McCool, Manuel Günther, and Sébastien Marcel. Bi-modal biometric authentication on mobile phones in challenging conditions. *Image and Vision Computing*, 32(12).

[95] Mohammad Rafiqul Alam, Roberto Togneri, Ferdous Sohel, Mohammed Bennamoun, and Imran Naseem. Linear regression-based classifier for audio visual person identification. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–5. IEEE, 2013.

[96] Qian Shi, Takeshi Nishino, and Yoshinobu Kajikawa. Multimodal person authentication system using features of utterance. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–7. IEEE, 2013.

[97] Roberto Brunelli and Daniele Falavigna. Person identification using multiple cues. *IEEE transactions on pattern analysis and machine intelligence*, 17(10):955–966, 1995.

[98] B. Maison, C. Neti, and A. Senior. Audio-visual speaker recognition for video broadcast news: some fusion techniques. In *1999 IEEE Third Workshop on Multimedia Signal Processing (Cat. No.99TH8451)*, pages 161–167, 1999.

[99] Conrad Sanderson. The vidtimit database. Technical report, IDIAP, 2002.

[100] P. Tresadern, T. F. Cootes, N. Poh, P. Matejka, A. Hadid, C. Lévy, C. McCool, and S. Marcel. Mobile biometrics: Combined face and voice verification for a mobile platform. *IEEE Pervasive Computing*, 12(1):79–87, 2013.

[101] Petr Motlicek, Laurent El Shafey, Roy Wallace, Christopher McCool, and Sébastien Marcel. Bi-modal authentication in mobile environments using session variability modelling. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1100–1103. IEEE, 2012.

[102] Md Rabiul Islam and Md Abdus Sobhan. Bpn based likelihood ratio score fusion for audio-visual speaker identification in response to noise. *ISRN Artificial Intelligence*, 2014, 2014.

[103] Mohammad Rafiqul Alam, Mohammed Bennamoun, Roberto Togneri, and Ferdous Sohel. A deep neural network for audio-visual person recognition. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2015.

[104] Mohammad Rafiqul Alam, Mohammed Bennamoun, Roberto Togneri, and Ferdous Sohel. A joint deep boltzmann machine (jdbm) model for person identification using mobile phone data. *IEEE Transactions on Multimedia*, 19(2):317–326, 2016.

[105] Rudi Primorac, Roberto Togneri, Mohammed Bennamoun, and Ferdous Sohel. Audio-visual biometric recognition via joint sparse representations. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3031–3035. IEEE, 2016.

[106] Q Memon, Z AlKassim, E AlHassan, M Omer, and M Alsiddig. Audio-visual biometric authentication for secured access into personal devices. In *Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science*, pages 85–89, 2017.

[107] Mikhail Gofman, Narciso Sandico, Sinjini Mitra, Eryu Suo, Sadun Muhi, and Tyler Vu. Multimodal biometrics via discriminant correlation analysis on mobile devices. In *Proceedings of the International Conference on Security and Management (SAM)*, pages 174–181. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2018.

[108] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22(6):123–151, 2005.

[109] Gunawan YB, Riyanto Bambang, et al. Feature level fusion of speech and face image based person identification system. In *2010 Second International Conference on Computer Engineering and Applications*, volume 2, pages 221–225. IEEE, 2010.

[110] Marc Acheroy, C Beumier, Josef Bigün, Gérard Chollet, Benoît Duc, Stefan Fischer, Dominique Genoud, Philip Lockwood, Gilbert Maitre, Stéphane Pigeon, et al. Multi-modal person verification tools using speech and images. *Multimedia Applications, Services and Techniques (ECMAST 96)*, 1996.

[111] Benoît Duc, Elizabeth Saers Bigün, Josef Bigün, Gilbert Maître, and Stefan Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):835–843, 1997.

[112] Souheil Ben-Yacoub. Multi-modal data fusion for person authentication using svm. Technical report, IDIAP, 1998.

[113] Conrad Sanderson and Kuldip K Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.

[114] R Herpers, G Verghese, K Derpanis, R McCready, J MacLean, A Levin, D Topalovic, L Wood, A Jepson, and JK Tsotsos. Detection and tracking of faces in real environments. In *Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No. PR00378)*, pages 96–104. IEEE, 1999.

[115] Linlin Shen, Nengheng Zheng, Songhao Zheng, and Wei Li. Secure mobile services by face and speech based personal authentication. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, volume 3, pages 97–100. IEEE, 2010.

[116] Luhong Liang, Xiaoxing Liu, Yibao Zhao, Xiaobo Pi, and Ara V Nefian. Speaker independent audio-visual continuous speech recognition. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 2, pages 25–28. IEEE, 2002.

[117] Ingrid Visentini, Christian Micheloni, and Gian Luca Foresti. Tuning asymboost cascades improves face detection. In *2007 IEEE International Conference on Image Processing*, volume 4, pages IV–477. IEEE, 2007.

[118] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[119] Danian Zheng, Yannan Zhao, and Jiaxin Wang. Features extraction using a gabor filter family. In *Proceedings of the sixth Lasted International conference, Signal and Image processing, Hawaii*, 2004.

[120] Lin Huang, Hanqi Zhuang, Sal Morgera, and Wenjing Zhang. Multiresolution pyramidal gabor-eigenface algorithm for face recognition. In *Third International Conference on Image and Graphics (ICIG'04)*, pages 266–269. IEEE, 2004.

[121] Kazuhiro Fukui and Osamu Yamaguchi. Facial feature point extraction method based on combination of shape extraction and pattern matching. *Systems and Computers in Japan*, 29(6):49–58, 1998.

[122] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650, 2010.

[123] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. *Computer vision using local binary patterns*, volume 40. Springer Science & Business Media, 2011.

[124] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005.

[125] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[126] Josef Kittler, YP Li, Jiri Matas, and MU Ramos Sánchez. Combining evidence in multimodal personal identity recognition systems. In *International*

Conference on Audio-and Video-based Biometric Person Authentication, pages 327–334. Springer, 1997.

[127] Frédéric Bimbot, Ivan Magrin-Chagnolleau, and Luc Mathan. Second-order statistical measures for text-independent speaker identification. *Speech communication*, 17(1-2):177–192, 1995.

[128] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

[129] Mohammad Rafiqul Alam, Mohammed Bennamoun, Roberto Togneri, and Ferdous Sohel. An efficient reliability estimation technique for audio-visual person identification. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pages 1631–1635. IEEE, 2013.

[130] Robbie Vogt and Sridha Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 22(1):17–38, 2008.

[131] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[132] Andrew O Hatch, Sachin Kajarekar, and Andreas Stolcke. Within-class covariance normalization for svm-based speaker recognition. In *Ninth international conference on spoken language processing*, 2006.

[133] Simon JD Prince and James H Elder. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[134] Chris McCool and Sébastien Marcel. Mobio database for the icpr 2010 face and speech competition. Technical report, Idiap, 2009.

[135] Roy Wallace, Mitchell McLaren, Christopher McCool, and Sébastien Marcel. Inter-session variability modelling and joint factor analysis for face authentication. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2011.

[136] Yongtao Hu, Jimmy SJ Ren, Jingwen Dai, Chang Yuan, Li Xu, and Wenping Wang. Deep multimodal speaker naming. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1107–1110. ACM, 2015.

[137] Grigory Antipov, Nicolas Gengembre, Olivier Le Blouch, and Gaël Le Lan. Automatic quality assessment for audio-visual verification systems. the love submission to nist sre challenge 2019. *arXiv preprint arXiv:2008.05889*, 2020.

[138] CC Chibelushi, F Deravi, and JS Mason. Bt david database-internal report. *Speech and Image Processing Research Group, Dept. of Electrical and Electronic Engineering, University of Wales Swansea [URL: http://wwwee. swan. ac. uk/SIPL/david/survey. html]*, 1996.

[139] Alberto Battocchi, Fabio Pianesi, and Dina Goren-Bar. Dafex: Database of facial expressions. In *International Conference on Intelligent Technologies for Interactive Entertainment*, pages 303–306. Springer, 2005.

[140] Stéphane Pigeon and Luc Vandendorpe. The m2vts multimodal face database (release 1.00). In Josef Bigün, Gérard Chollet, and Gunilla Borgefors, editors, *Audio- and Video-based Biometric Person Authentication*, pages 403–409, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[141] Girija Chetty. Biometric liveness detection based on cross modal fusion. In *2009 12th International Conference on Information Fusion*, pages 2255–2262. IEEE, 2009.

[142] Zheng-Yu Zhu, Qian-Hua He, Xiao-Hui Feng, Yan-Xiong Li, and Zhi-Feng Wang. Liveness detection using time drift between lip movement and voice. In *2013 International Conference on Machine Learning and Cybernetics*, volume 2, pages 973–978. IEEE, 2013.

[143] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. Avicar: Audio-visual speech corpus in a car environment. In *Eighth International Conference on Spoken Language Processing*, 2004.

[144] A. F. Sequeira, J. C. Monteiro, A. Rebelo, and H. P. Oliveira. Mobbio: A multimodal database captured with a portable handheld device. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, pages 133–139, Jan 2014.

[145] Hervé Bredin and Gérard Chollet. Making talking-face authentication robust to deliberate imposture. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1693–1696. IEEE, 2008.

[146] Gérard Chollet, Rémi Landais, Thomas Hueber, Hervé Bredin, Chafic Mokbel, Patrick Perrot, and Leila Zouari. Some experiments in audio-visual

speech processing. In *International Conference on Nonlinear Speech Processing*, pages 28–56. Springer, 2007.

[147] P. W. McOwan and A. Johnston. The algorithms of natural vision: the multi-channel gradient model. In *First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications*, pages 319–324, 1995.

[148] Nicolas Eveno, Alice Caplier, and P-Y Coulon. Accurate and quasi-automatic lip tracking. *IEEE Transactions on circuits and systems for video technology*, 14(5):706–715, 2004.

[149] Nicolas Eveno and Laurent Besacier. Co-inertia analysis for" liveness" test in audio-visual biometrics. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 257–261. IEEE, 2005.

[150] Jorge Gutierrez, J-L Rouas, and Régine André-Obrecht. Weighted loss functions to make risk-based language identification fused decisions. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 863–866. IEEE, 2004.

[151] N Mana, P Cosi, G Tisato, F Cavicchio, E Caldognetto Magno, and F Pianesi. An italian database of emotional speech and facial expressions. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006*, page 68, 2006.

[152] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.

[153] Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. Multilingual voice impersonation dataset and evaluation. 2020.

[154] Hareesh Mandalapu, Thomas Møller Elbo, Raghavendra Ramachandra, and Christoph Busch. Cross-lingual speaker verification: Evaluation on x-vector method. In Sule Yildirim Yayilgan, Imran Sarwar Bajwa, and Filippo Sanfilippo, editors, *Intelligent Technologies and Applications*, pages 215–226, Cham, 2021. Springer International Publishing.

[155] Ashok Kumar Das, Mohammad Wazid, Neeraj Kumar, Athanasios V Vasilakos, and Joel JPC Rodrigues. Biometrics-based privacy-preserving user authentication scheme for cloud-based industrial internet of things deployment. *IEEE Internet of Things Journal*, 5(6):4900–4913, 2018.

[156] Vishal M Patel, Nalini K Ratha, and Rama Chellappa. Cancelable biometrics: A review. *IEEE Signal Processing Magazine*, 32(5):54–65, 2015.

[157] Patrizio Campisi. *Security and privacy in biometrics*, volume 24. Springer, 2013.

[158] Mehrdad Aliasgari and Marina Blanton. Secure computation of hidden markov models. In *2013 International Conference on Security and Cryptography (SECRYPT)*, pages 1–12. IEEE, 2013.

[159] Nitin Kumar et al. Cancelable biometrics: a comprehensive survey. *Artificial Intelligence Review*, pages 1–44, 2019.

[160] Margarita Osadchy, Benny Pinkas, Ayman Jarrous, and Boaz Moskovich. Scifi-a system for secure face identification. In *2010 IEEE Symposium on Security and Privacy*, pages 239–254. IEEE, 2010.

[161] Ahmad-Reza Sadeghi, Thomas Schneider, and Immo Wehrenberg. Efficient privacy-preserving face recognition. In *International Conference on Information Security and Cryptology*, pages 229–244. Springer, 2009.

[162] Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Lagendijk, and Tomas Toft. Privacy-preserving face recognition. In *International symposium on privacy enhancing technologies symposium*, pages 235–253. Springer, 2009.

[163] Andrew Beng Jin Teoh and Lee-Ying Chong. Secure speech template protection in speaker verification system. *Speech communication*, 52(2):150–163, 2010.

[164] Wenhua Xu and Minying Cheng. Cancelable voiceprint template based on chaff-points-mixture method. In *2008 International Conference on Computational Intelligence and Security*, volume 2, pages 263–266. IEEE, 2008.

[165] Marco Paulini, Christian Rathgeb, Andreas Nautsch, Hermine Reichau, Herbert Reininger, and Christoph Busch. Multi-bit allocation: Preparing voice biometrics for template protection. In *Odyssey*, pages 291–296, 2016.

[166] Stefan Billeb, Christian Rathgeb, Herbert Reininger, Klaus Kasper, and Christoph Busch. Biometric template protection for speaker recognition based on universal background models. *IET Biometrics*, 4(2):116–126, 2015.

[167] Aymen Mtibaa, Dijana Petrovska-Delacrétaz, and Ahmed Ben Hamida. Cancelable speaker verification system based on binary gaussian mixtures. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6. IEEE, 2018.

[168] Andreas Nautsch, Sergey Isadskiy, Jascha Kolberg, Marta Gomez-Barrero, and Christoph Busch. Homomorphic encryption for speaker recognition: Protection of biometric templates and vendor model parameters. *arXiv preprint arXiv:1803.03559*, 2018.

[169] Amos Treiber, Andreas Nautsch, Jascha Kolberg, Thomas Schneider, and Christoph Busch. Privacy-preserving plda speaker verification using outsourced secure computation. *Speech Communication*, 114:60–71, 2019.

[170] Slobodan Ribaric and Nikola Pavesic. An overview of face de-identification in still images and videos. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 4, pages 1–6. IEEE, 2015.

[171] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9378–9387, 2019.

[172] Manas A Pathak and Bhiksha Raj. Privacy-preserving speaker verification as password matching. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1849–1852. IEEE, 2012.

[173] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–1. IEEE, 2004.

[174] Patrick J Grother, Patrick J Grother, and Mei Ngan. *Face recognition vendor test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology, 2014.

[175] P Grother, GW Quinn, JR Matey, M Ngan, W Salamon, G Fiumara, and C Watson. Irex-iii: Performance of iris identification algorithms. nist interagency report 7836. *NIST, Gaithersburg, MD*, 2012.

[176] Craig I Watson, Gregory P Fiumara, Elham Tabassi, Su L Cheng, Patricia A Flanagan, and Wayne J Salamon. Fingerprint vendor technology evaluation. Technical report, 2015.

[177] Ajita Rattani and Reza Derakhshani. A survey of mobile face biometrics. *Computers & Electrical Engineering*, 72:39–52, 2018.

[178] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[179] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[180] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[181] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019.

[182] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[183] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[184] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[185] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[186] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[187] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[188] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[189] Simon Prince, Peng Li, Yun Fu, Umar Mohammed, and James Elder. Probabilistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157, 2011.

[190] Nam Le and Jean-Marc Odobez. Robust and discriminative speaker embedding via intra-class distance variance regularization. In *Interspeech*, pages 2257–2261, 2018.

[191] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[192] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. volume 1, pages 41.1–41.12. British Machine Vision Association, 01 2015.

[193] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[194] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. 2015.

[195] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey*, volume 2016, pages 283–290, 2016.

[196] Ivana Chingovska, Andre Rabello Dos Anjos, and Sebastien Marcel. Biometrics evaluation under spoofing attacks. *IEEE transactions on Information Forensics and Security*, 9(12):2264–2276, 2014.

[197] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015.

[198] A. Misra and J. H. L. Hansen. Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 372–377, 2014.

[199] Lantian Li, Dong Wang, Askar Rozi, and Thomas Fang Zheng. Cross-lingual speaker verification with deep feature learning. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1040–1044. IEEE, 2017.

[200] Ahmad N Al-Raisi and Ali M Al-Khouri. Iris recognition and the challenge of homeland and border control security in uae. *Telematics and Informatics*, 25(2):117–132, 2008.

[201] Diego Gragnaniello, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Contact lens detection and classification in iris images through scale invariant descriptor. In *2014 Tenth International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 560–565. IEEE, 2014.

[202] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Generalized textured contact lens detection by extracting bsif description from cartesian iris images. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–7. IEEE, 2014.

[203] Ramachandra Raghavendra, Kiran B Raja, and Christoph Busch. Ensemble of statistically independent filters for robust contact lens detection in iris images. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 24. ACM, 2014.

[204] James S Doyle and Kevin W Bowyer. Robust detection of textured contact lenses in iris recognition using bsif. *IEEE Access*, 3:1672–1683, 2015.

[205] Naman Kohli, Daksha Yadav, Mayank Vatsa, Richa Singh, and Afzel Noore. Detecting medley of iris spoofing attacks using desist. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–6. IEEE, 2016.

[206] Yang Hu, Konstantinos Sirlantzis, and Gareth Howells. Iris liveness detection using regional features. *Pattern Recognition Letters*, 82:242–250, 2016.

[207] Zhuoshi Wei, Xianchao Qiu, Zhenan Sun, and Tieniu Tan. Counterfeit iris detection based on texture analysis. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[208] Zhaofeng He, Zhenan Sun, Tieniu Tan, and Zhuoshi Wei. Efficient iris spoof detection via boosted local binary patterns. In *International Conference on Biometrics*, pages 1080–1090. Springer, 2009.

[209] Hui Zhang, Zhenan Sun, and Tieniu Tan. Contact lens detection based on weighted lbp. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4279–4282. IEEE, 2010.

[210] Zhenan Sun, Hui Zhang, Tieniu Tan, and Jianyu Wang. Iris image classification based on hierarchical visual codebook. *IEEE Transactions on pattern analysis and machine intelligence*, 36(6):1120–1133, 2014.

[211] Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1363–1366. IEEE, 2012.

[212] Kiran B Raja, Ramachandra Raghavendra, and Christoph Busch. Binarized statistical features for improved iris and periocular recognition in visible spectrum. In *Biometrics and Forensics (IWBF), 2014 International Workshop on*, pages 1–6. IEEE, 2014.

[213] Kiran B Raja, Ramachandra Raghavendra, and Christoph Busch. Color adaptive quantized patterns for presentation attack detection in ocular biometric systems. In *Proceedings of the 9th International Conference on Security of Information and Networks*, pages 9–15. ACM, 2016.

[214] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 29794-6 Information Technology - Biometric Sample Quality - Part 6: Iris Image Data*. International Organization for Standardization, 2015.

[215] Anna Bori Toth and Javier Galbally. Anti-spoofing, iris. *Encyclopedia of Biometrics*, pages 87–97, 2015.

[216] Stan Z Li and Anil K Jain. *Encyclopedia of Biometrics: I-Z.*, volume 1. Springer Science & Business Media, 2009.

[217] Hong Wei, Lulu Chen, and James Ferryman. Biometrics in abc: counterspoofing research. 2013.

[218] Kevin W Bowyer and James S Doyle. Cosmetic contact lenses and iris recognition spoofing. *Computer*, 47(5):96–98, 2014.

[219] Ramachandra Raghavendra, Kiran B Raja, and Christoph Busch. Contlensnet: Robust iris contact lens detection using deep convolutional neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1160–1167. IEEE, 2017.

[220] Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. Image quality and texture-based features for reliable textured contact lens detection (article in press). In *Signal Imaging Technology and Internet Based Systems, 2018. SITIS 2018. 14th International Conference on*. IEEE, 2018.

[221] Sushma Venkatesh, R Raghavendra, Kiran B. Raja, and Christoph Busch. A new multi-spectral iris acquisition sensor for biometric verification and presentation attack detection. 11 2018.

[222] Juefei Xu, Miriam Cha, Joseph L Heyman, Shreyas Venugopalan, Ramzi Abiantun, and Marios Savvides. Robust local binary pattern feature sets for periocular biometric identification. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8. IEEE, 2010.

[223] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.

[224] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66:130–153, 2015.

[225] Mireia Farrús Cabeceran, Michael Wagner, Daniel Erro Eslava, and Francisco Javier Hernando Pericás. Automatic speaker recognition as a measurement of voice imitation and conversion. *The Intenational Journal of Speech. Language and the Law*, 1(17):119–142, 2010.

[226] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, 72:13–31, 2015.

[227] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[228] Yee W Lau, Dat Tran, and Michael Wagner. Testing voice mimicry with the yoho speaker verification corpus. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 15–21. Springer, 2005.

[229] Yee Wah Lau, M. Wagner, and D. Tran. Vulnerability of speaker verification to voice mimicking. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pages 145–148, Oct 2004.

[230] Johnny Mariéthoz and Samy Bengio. Can a professional imitator fool a gmm-based speaker verification system? Technical report, IDIAP, 2005.

[231] Saurabh Panjwani and Achintya Prakash. Crowdsourcing attacks on biometric systems. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 257–269, 2014.

[232] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017.

[233] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2616–2620. ISCA, 2017.

[234] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, March 2006.

[235] Matthieu Hébert. Text-dependent speaker recognition. In *Springer handbook of speech processing*, pages 743–762. Springer, 2008.

[236] Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83, 1995.

[237] K. S. R. Murty and B. Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE Signal Processing Letters*, 13(1):52–55, 2006.

[238] Cheedella S Gupta. Significance of source features for speaker recognition. *Master's thesis, Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India*, 2003.

[239] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5):569–586, 1999.

[240] J. Wang and M. T. Johnson. Vocal source features for bilingual speaker identification. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 170–173, 2013.

[241] Artur Janicki, Federico Alegre, and Nicholas Evans. An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. *Security and Communication Networks*, 9(15):3030–3044, 2016.

[242] Héctor Delgado, Massimiliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, and Junichi Yamagishi. Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements. In *Odyssey 2018-The Speaker and Language Recognition Workshop*, 2018.

[243] Hemant A Patil and Madhu R Kamble. A survey on replay attack detection for automatic speaker verification (asv) system. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1047–1053. IEEE, 2018.

[244] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2017. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, March 2017.

Part III

Appendix

Chapter 13

Appendix A

The MAVS data capturing application and details are explained in this section.

13.1 Mobile Application

The process of data capturing is carried out with the help of a mobile application implemented on iOS and Android platforms. The application is a modified version of the SWAN data capturing mobile application. The application is distributed as a compiled file for iOS (.ipa) and Android (.apk) environments. The mobile application is operated by data capturing subjects. Therefore, the application is made easy and user-friendly.

13.1.1 Data Storage

The biometric data files are stored internally right after each recording. The file names are created as shown below.

Filename: `subjectid_devicename_session_langauge_textid.mp4`

Example: `0910M_samsungs8_session1_english_4.mp4`

The files can be transferred to other devices from the file storage of the mobile devices.

13.2 Capturing GUI

The application starts with an information page to input the user details. The application opens the front camera with instructions to start the recording as "RECORD". A text is prompted on the top of the display for the subject to read during the recording. The subject should stop the recording by pressing the "STOP" button when the recording is finished. The recorded file is saved with a subject-

specific file name right after the capturing. Further, the second sentence appears on the screen, and the subject continues to record all sentences in different languages. Figure 13.1 shows the interface of the android application for data capturing. The iOS application is also made to appear as similar as possible for user-friendliness.

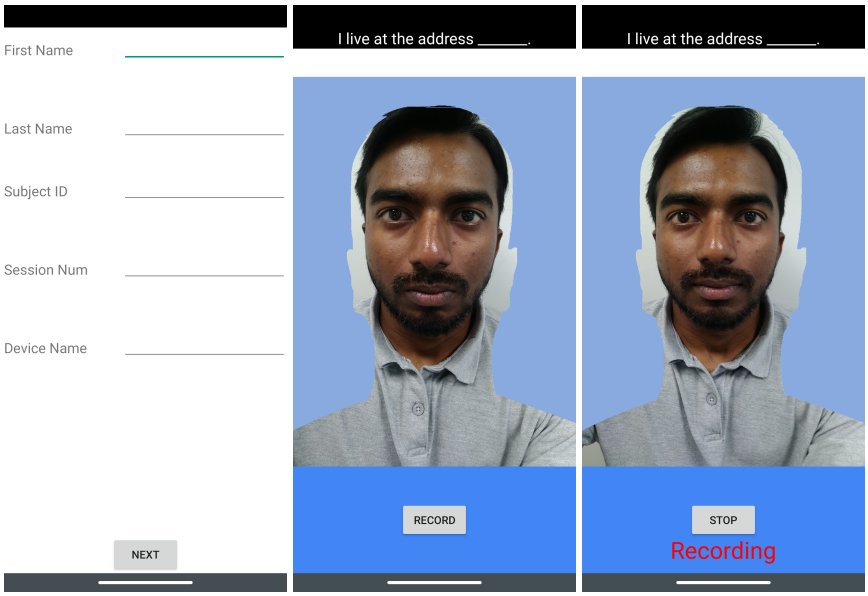


Figure 13.1: Android application interface

ISBN 978-82-326-6792-5 (printed ver.)
ISBN 978-82-326-5792-6 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology