Multimodal Authentication Systems:

A Consideration of System Integrity, Availability and Resilience Against Spoofing Attacks

by

Aitore Issadykova and Assem Kussainova

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Data Science

at the

NAZARBAYEV UNIVERSITY

April 2021

Author .....
Aitore Issadykova
April 30, 2021

Author .....
Assem Kussainova
April 30, 2021

Certified by .....
Michael Lewis
Associate Professor
Thesis Supervisor

Accepted by .....
Vassilios D. Tourassis
Dean, School of Engineering and Digital Sciences

Multimodal Authentication Systems: A Consideration of System Integrity, Availability and Resilience Against Spoofing Attacks

by

Aitore Issadykova and Assem Kussainova

Submitted to the Department of Computer Science
on April 30, 2021, in partial fulfillment of the
requirements for the degree of
Master of Data Science

Abstract

User authentication is a fundamental requirement of any role-based access control system, governing both physical and digital access to organizational resources, and the related security and privacy of data and transactional meta-data. In our paper we review methods of authentication based on biometric characteristics such as fingerprint, retina, hand geometry, face geometry, face thermogram, voice and handwriting. We replicated recent work on multimodal biometric authentication, using aligned streams of audio and video data, and examined obfuscation techniques that could be used to undermine confidence in those techniques.

Based on this experience, we designed and implemented a system for combined face and voice authentication using the open-access SpeakingFaces dataset. Vocal features are extracted using Mel-Frequency Cepstral Coefficients (MFCCs), and facial features are obtained with Local Binary Patterns (LBPs). Face and voice identification are performed using image similarity with the Euclidean distances metric and Gaussian Mixture Model (GMM) respectively, and in turn combined into a single multimodal system using matching scores fusion.

The multimodal biometric authentication system was assessed using open-source data from Georgia Tech Face Database and the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. The confidentiality of the face and voice recognition system was analyzed with several scenarios using spoofing of facial features, imitation of voice features, combined spoofing, and no spoofing scenarios.This project successfully replicated the published work, improved the computational performance and demonstrated that the ranking based model of multimodal biometric system is more resilient than a threshold-based system. Reported weaknesses of the prior works were used to improve the performance of our multimodal biometric authentication system.

Thesis Supervisor: Michael Lewis
Title: Associate Professor

Acknowledgments

During the completion of practical and writing parts of this thesis we have received a great deal of support and assistance.

We would first like to thank our lead advisor, Professor Michael Lewis, whose advises were invaluable in selecting research area. Your insightful feedbacks and weekly report meetings pushed us to sharpen our presentation and writing skills and brought our work to a higher level.

We would also like to thank our co-advisor, Dr. Dimitrios Zorbas, for his dedication to the project by allocating personal time for proof reading and advising of the thesis work.

We would like to acknowledge Dr. Fabian Monrose for his willingness to be an external supervisor for this work. We would particularly like to thank you for your valuable comments about projects and advises for its improvement. We honored to have an expert in the biometric authentication field as our advisor and we honored to have an opportunity for thesis revision by you.

We would also like to highlight the role of our Dean, Professor Vassilios D.Tourassis in our project. We would like to thank you in your support for our joined thesis work and provide valuable instruction toward its completion.

In addition, We would like to thank ourselves for continuous support of each other and understanding. Finally, We would like to sincerely thank our friends from Master Program in Data Science, who provided happy distractions to rest our minds outside of our research.

Contents

Specific annotations for work done individually: *(AI)* - Aitore Issadykova, *(AK)* - Assem Kussainova - otherwise, the contribution is collaborative

Chapter 1

Introduction

The aim of this thesis is to recreate and extend a multimodal biometric authentication system based on recently published works, and then examine the confidentiality, integrity and availability of the system using two open-source datasets and the novel new large-scale SpeakingFaces dataset.

The work is organized into eight sections.

Chapter One introduces the motivation behind using biometric techniques to enhance the security of systems and the integrity of data, and reviews the strengths and weaknesses of multimodal biometric authentication systems. We present the objectives of this thesis, and describe the phases and procedures of the implementation of our system.

Chapter Two describes the current state-of-the-art based on a review of the relevant literature, summarizing the contributions of other researchers in the areas of unimodal and multimodal biometric systems, and the performance of their systems. We identify recent published examples which were then used as a foundation for the current work, along with open-source datasets used for the purpose of benchmarking.

Chapter Three provides a detailed description of the datasets used in this work: the TIMIT dataset assembled by a collaboration of MIT, SRI International, and Texas Instruments; the Georgia Tech Face Database; and the SpeakingFaces dataset of the Institute for Smart Systems and Artifical Intelligence (ISSAI) of Nazarbayev University. We describe their respective methods of collection, data attributes, limitations,

and the preprocessing necessary to use the data.

Chapter Four presents the architecture of the multimodal biometic system that we have chosen to replicate. The authentication system contains a facial feature recognition component, a voice pattern extraction component and a fusion score to produce the multimodal system from two unimodal segments. All segments provide a detailed description of the whole intrinsic model, with the specific templates and parameter setup for the subset of data used from the SpeakingFaces Dataset.

Chapter Five presents the results achieved from running our system on the subset of SpeakingFaces. This chapter includes discussions of the impact of each singular parameter of the system on the overall performance, and cross validation of the performance of the recreated biometric authentication system. We assess the accuracy of the system on the acknowledged datasets such as TIMIT and the Georgia Tech Face Database. The achievements of other researchers are compared for each unimodal component respectively.

Chapter Six introduces the concept of spoofing techniques, explicitly designed to obfuscate the biometric measures, and their impact on the security level of the authentication system. Four different scenarios are considered: no spoofing, spoofing on the face features only, spoofing of the voice pattern and combined spoofing. We ran each of the scenarios in the replicated multimodal system, and prepared an assessment of the impact of spoofing on the authentication process. In addition, for the combined spoofing scenario we created synthetic biometric data using morphed face images and voice imitation audio.

Chapter Seven examines all outcomes of the previous chapters, and takes into account the known weaknesses of the prior works. We then describe the refinements we designed and implemented two fusion scores, and compared their relative performance. In addition, the comprehensive evaluation and comparison of the system integrity and availability with time-cost analysis is reported.

Chapter Eight presents all of the achievements obtained on the multimodal biometric authentication system. The goals and aims of the thesis are analyzed, with observations regarding the potential weaknesses of the work, and concludes with a

consideration of possible future work.

1.1 Motivation Behind Using Biometric Data

The use of information systems has become an integral part of most spheres of human activity. Access to these systems and the security of data stored there was most often controlled through the allocation of an online identity, represented by a User ID and a password. However, the combination of User ID and password has proven an ineffective means of user authentication, indeed, it does not serve to reliably connect an online identity with an actual person.

The need for reliable user authentication as a means of access control motivated the development of multi-factor authentication, typically consisting of three components: something you know, something you have, and something that you are.

This third factor of "something that you are" is most often implemented through user biometrics, as they are considered relatively unique and not easily replicated by intruders.

1.1.1 Historical Overview

The first computerized databases and related data storage systems were introduced in the 1960s; prior to this point, most documents and data were stored in the format of paper, which, though reasonably secure and private, by way of limited access, were for the same reason not readily consulted, nor easily searched.

The gradual development of computer technologies and digital storage media soon demonstrated the potential advantages of the new systems, in terms of reduced storage requirements, and enabling the indexing of contents, thereby facilitating rapid search based on keyword combinations.

The improvement of hardware and internet technologies led to the generation of enormous volumes of textual data, but also extending to include multimedia data such as high-resolution audio and video. The volume of the data currently stored surpasses the mark of the 59 Zetabytes, and this indicator is expected to rise exponentially in

the coming decade.

The acquisition, storage, and analysis of data often overwhelms the ability to store the data on a local machine or server, which in turn has prompted the emergence of the concept of Cloud storage, and Cloud-based computation. However, the rapid shift from localized storage to Cloud-based storage served to emphasize the concerns for the security and privacy of data stored remotely, outside the direct physical control of the user.

The limitations of traditional user authentication have been further exposed with many high-profile cases of large-scale data breaches and compromises. In order to improve the security of systems, and the privacy and security of data storage, it has become imperative to consider the use of multi-factor user authentication that incorporates biometric data.

1.1.2 Biometric Authentication Systems

The utility of biometric authentication was recognized from the early stages within the cyber security community, which engaged in proof-of-concept experimentation with a wide range of biometric data sources such as fingerprints, retina scans, facial features, and voice patterns as well as behavioral biometrics such as keystroke dynamics, voice tempo and pronunciation examination, analysis of mouse use and touch pad dynamics, along with cognitive biometrics. All can be used, either individually or in combinations, to create new modern technologies designed to control access to system resources, and protect the privacy and security of data stored there.

There is considerable debate as to which type of biometric data are best suited, nevertheless, the majority of the innovations used physical biometric data. Generally, physical systems that control access to facilities or devices utilize approaches such as facial recognition, fingerprints, and retina scans, while online systems can also utilize the behavioral biometric variants. However, such systems remain vulnerable, as they can inadvertently reject legitimate users through too-strict of implementation, yet remain vulnerable to replication or spoofing as a means to undermine the reliability of the systems.

In our work, we review recent work regarding biometric authentication systems, examining several approaches by themselves and in combination, and then conduct an analysis in terms of robustness and overall integrity of the systems.

1.2 Thesis Motivation

User authentication is a fundamental requirement of any role-based access control system, governing both physical and digital access to organizational resources and sensitive data. The real-world objective is to provide methods for user authentication that are resistant to spoofing and subversion. As noted, many hybrid and multimodal biometric recognition techniques have been presented to control system access and protect data, incorporating both physical and behavioural biometric techniques. The multimodal biometric systems were introduced to reflect the issues and limitations of both traditional password-based and unimodal biometric systems of authentication.

The majority of systems and related research focus on the improvement of the authentication of the real user by increasing the True Acceptance Rate (TAR) of the system and diminishing the False Rejection Rate (FRR), without imposing undue obstacles to the user or delays in accessing system resources. As an additional improvement of the security of biometric authentication, "liveness" detection was introduced as a means to reduce system vulnerability.

Nevertheless, the TAR and FRR do not provide sufficient assurance of the security level of the system. Such systems remain vulnerable to attacks on the multimodal authentication system and spoofing of the biometric data by intruders . Taking into consideration the high volumes of compromised personal information and spoofed biometric data, many system developers and researchers do not provide adequate consideration of the performance of the authentication methods while under stress or attack. In some cases, the methods of spoofing used to demonstrate system robustness are considered outdated, and not indicative of current intruder methods.

Recent innovations in the fabrication of biometric data, such as face morphing and

voice imitation using Hidden Markov Models, serve to illustrate potential system vulnerabilities. Such approaches should also be used to analyze the possible weaknesses of the new authentication models, thus providing improved scrutiny of the security, availability and integrity of the system.

1.3 Goals and Aims of the Current Work

The Internet is a resilient store-and-forward network, of historic importance. But security was not a design objective. Similarly, the early generations of many popular systems and applications were driven by considerations of first-to-market, often relinquishing security to a lower priority.

The introduction of multi-factor authentication to control access to systems and resources is welcome, but it, too, has emerged in stages, with early approaches achieving incremental gains but remaining vulnerable.

We emphasize in our work the utility of multimodal over unimodal biometrics, as the multimodal approach can compensate for the inherent limitations of unimodal systems. Further, we advocate for consideration of adversarial spoofing during development and deployment as a means to improve overall security and robustness of the system.

In our work, we replicate recent work in the field, using two open-source multimodal datasets, and a new one that includes audio, video, and thermal data streams. We identify potential security problems by analysing the behaviour of the system under spoofing attacks.

For the validation purposes, we assess the performance of each unimodal authentication component with the well-known TIMIT dataset and Georgia Tech Face database, and the results are compared with the achievements of the original manuscript using the same authentication model, and with the performance of other works.

Based on the replication and performance analysis, we designed a variation of the system and introduce a fusion scoring system comprised of face recognition and voice pattern extraction components on the subset of the newly collected SpeakingFaces

dataset.

This paper also critically examines the effectiveness of the multimodal biometric system for data protection purposes. The scrutiny of the security of the authentication model is analyzed under several different configurations of spoofing attacks. In this investigation, the aim is to assess the reliability of the multimodal biometric system, and to generate a report on the potential weaknesses of any multimodal biometric system.

In the pages that follow, we review the performance of the multimodal authentication system and suggest possible improvements both to the model and the hardware components in order to minimize the probability of the False Acceptance Rate (FAR) for an intruder, and escalate the True Rejection Rate (TRR) of the system. Moreover, we improve the availability of the system by making enhancements based on the provided time-cost analysis. The integrity of the whole structure is evaluated as the last step, after implementing the system improvements. To sum up all project goals, we implement a multimodal biometric system and investigate system performance in terms of recognition rates, computational resources (time and space) and the resilience of the system to a range of attacking scenarios that utilize high-quality synthetic biometric data. As a conclusion, all procedures of the current work are assessed within the scope of the framework, with known limitations mentioned, along with recommendations for potential future work.

Chapter 2

Literature Review

Recently, an overwhelming amount of research has been done in the field of biometric authentication using physical and behavioural biometric characteristics. Automatic user recognition for identification has a large number of security and privacy applications in scenarios that involve authorizing access to controlled physical or digital resources. Though development of biometric recognition methods has been underway for several decades, the problem is still far from solved. However, there has been significant progress recently, due to the development of new methods, specialized computational devices, and the emergence of large databases of biometric information suitable for training of data-hungry machine-learning systems. As a result, biometric recognition systems are rapidly emerging at the practical level for purposes of security and access control to devices and data.

The most common biometric authentication methods have been face recognition, fingerprint and eye retina matching. These methods are frequently used in systems that control access to physical facilities and digital devices, and verification of traveler identity in scenarios such as border control.

This chapter describes recent studies of biometric recognition systems, covering both unimodal components and multimodal architectures. The major focus is given for physical and behavioral biometric authentication methods such as facial recognition and voice identification. We summarize other physical methods, as well as examples of behavioral biometrics, and introduce a brief overview of spoofing meth-

ods and how they might be used to overcome biometric authentication methods.

2.1 Unimodal Biometric Components

As noted, research in biometric methods and related data collection and analysis has become trendy, resulting in the rapid development of tools and techniques used for purposes of authentication. The foundation of this work begins with unimodal biometric components, that is, the data and methods specific to each biometric feature such as eyes, face, or voice.

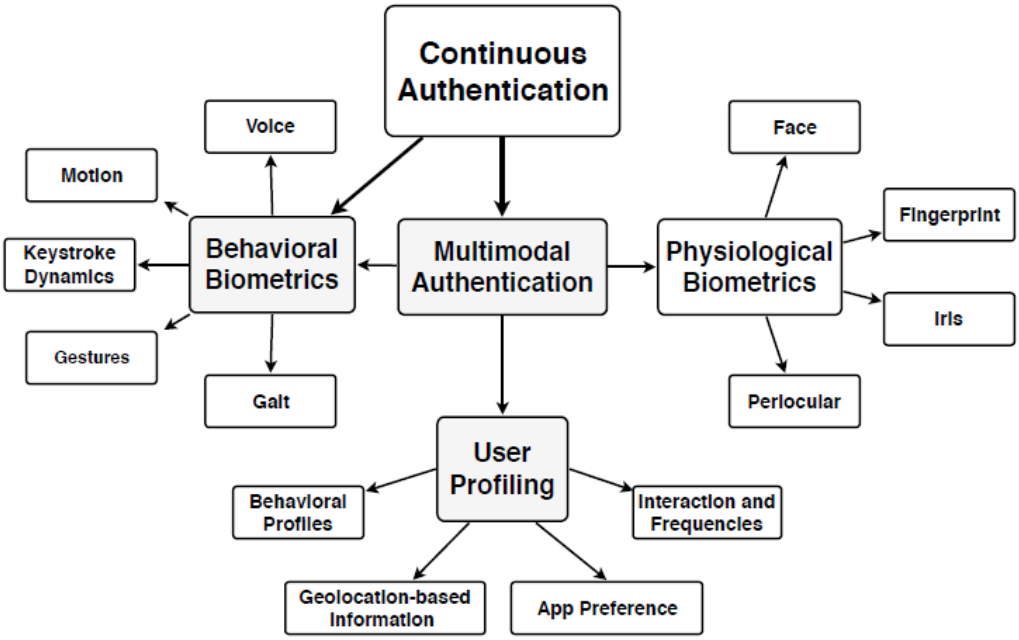


Figure 2-1: Biometric based authentication methods categories [3]

Fig. 2-1 illustrates the diversified modalities of the biometric authentication methods [3]. The major categories of unimodal systems are physiological biometrics, based on distinctiveness of user features, and behavioural biometrics, based on patterns of use such as keystroke dynamics or drawing a shape on a touch pad. Interest in behavioral biometrics are partially motivated by the concern that physiological biometrics did not provide a sufficient level of authentication, though it is also noted that behavioural biometrics are vulnerable to environmental elements such as temperature

or the physical and emotional state of the user.

2.1.1 Facial Features Recognition *(AI)*

Facial recognition systems have become a conventional way to gain entrance to controlled physical premises, and to unlock smart-devices. While the standard RGB image is most frequently used, thermal image data is gaining interest, due to recent advances in the resolution of thermal sensors, as a way to augment the RGB recognition process to both improve recognition rates and increase the resilience against spoofing.

The majority of the literature is focused on the improvement of the existing facial recognition systems, and the rest devoted to the exploration of new approaches; a recent trend is to combine strategies, and generate a hybrid model. The work is often benchmarked using well-known datasets such as the Georgia Tech Face Database, FERET [39]. Studies [60] and [31] proposed an effective model for the Georgia Tech Face Database. Research conducted by Zhang et al.[60] examined performance of the combined Haar cascades model with Local Binary Pattern (LBP) operator, while Lu et al.[31] achieved better results with an improved Ternancy color LBP (TCLBP) operator combined with the Enhanced Fisher linear discriminant Model (EFM). Wavelet decomposition with Linear Regression Classifier (LRC) [38] and String Grammar Nearest Neighbor [26] approaches presented unique techniques which were implemented and validated on open-access datasets. Statistical approaches such as the Jacobi method based on principal component analysis (PCA) and linear discriminant analysis are still prevalent techniques for real-time face recognition [12].

Advances in the deep learning sphere also affected the recognition systems. Deep Convolutional Neural Networks (CNN) are used in both feature detection and feature extraction modules. CNN outperformed the classical methods such as LBP and PCA [40] and demonstrated exceptional results on the RGB-D images [23]. Advanced approaches such as transfer learning of the famous AlexNet CNN with Support Vector Machine (SVM) classifier presented the ideal result of 100% accuracy on the Georgia Tech Face Database [5]; though transfer learning may result in high data bias, and

the model will require improvement for real-time detection.

Researchers	**Scientific Approach**	**Achieved TAR**
Almabdy S. and Elrefaei L. [5]	Transfer learning of pre-trained AlexNet with SVM classifier	**100%**
Lu et al. [31]	Enhanced Fisher linear discriminant Model (EFM) with Ternary-Color LBP (TCLBP)	94.57%
Zhang et al. [60]	Haar cascades with LBP features extraction	90.14%
Nunes et al. [38]	Wavelet decomposition with Linear Regression Classification (LRC).	80.3%
Kasemsumran et.al [26]	String Grammar Nearest Neighbor	70.71%

Table 2.1: Achieved TAR on the whole Georgia Tech Face Database

Table 2.1 illustrates a current the state-of-art approach which was implemented on the Georgia Tech Face Database and the results achieved by their model.

The creation of the facial recognition system based on thermal images required substantial data collection of the different states and moods of the person as well as the additional temperature analysis based on the environmental conditions. The study led by Mate Krišto and Marina Ivašić-Kos [29] described the fundamentals and basic concepts of infrared imaging, as well as methods of thermal facial image recognition and their characteristics, and recommended using Deep Learning and CNN for a more effective result when working with thermal images. Some studies are still working on the improvement of the statistical approaches such as Optimized Probability Density Function (OPDF) using Histogram of Dynamic Thermal Patterns (HDTP) as a descriptor [22]. Despite the fact that neural networks are costly for multimodal systems, model approaches combined neural networks with common algorithms such as PCA [30] and Multi-Block LBP combined with Adaboost using various classifiers [32].

2.1.2 Voice-Based Authentication *(AK)*

Voice recognition is used for biometric security purposes to identify the voice of a specific person. The acoustic characteristics of an individual's voice, such as tone and speaking style, differ from person to person. In voice biometrics for authentication, the voice is digitized and compared with a previously recorded template. Depending on the principle of operation, voice recognition systems are divided into those working with a text pattern (the comparison is made with a sample of previously read text) and those working with a voice (the voice features are compared).

Audio based authentication systems depend on the existence of microphone sensors and thus are mostly implemented on devices that routinely have microphones, such as mobile phones and laptop computers. The majority of studies for voice biometric models cover the voice features, and vary by implementing different methods of feature extraction and model creation. Mel-Frequency Cepstral Coefficients (MFCC) is one of the most common methods used to extract audio features. Studies [61], [14] and [47] used MFCC and Gaussian Mixture Models method (GMM) for building unimodal voice biometric systems. Authors [61] and [14] verified the performance of the model on the well-known TIMIT dataset (Table. 2.2), while researchers [47] built a model to test the Russian Speaking Corpus and achieved the highest accuracy of 98%. Boles and Rad [11] proposed a machine learning model using MFCC feature extraction and a neural network with an SVM classifier and validated on publicly available LibriSpeech dataset with achieved 95% accuracy on the data of 10 speakers. Another study conducted by Hendryli J. and Herwindiati D. [20] also combined the MFCC feature extraction with a mixture of Long Short-Term Memory (LSTM) and Siamese networks, which uses two identical CNNs; the accuracy of the model on the data collected was 61% for 23 speakers.

Other researchers substituted the feature extraction methods as well as the model or classifier approach. Studies [4] and [37] created models with I-vector with three fusion methods and Multi-Factor Authentication method with Discrete Wavelet Packet Transform (DWPT) and Quantization Index Modulation (QIM) respectively and val-

idated the performance of the models on the TIMIT dataset (Table. 2.2). While developing machine learning methods many studies [52], [41], [20] started to implement them as classifiers. The study conducted by Thullier et al. [52] replaced MFCC with Linear Prediction Cepstral Coefficients and used Naive Bayes classication for the voice authentication model, and compared results on the TIMIT dataset (Table. 2.2). Chowdhury et al. [41] used "OK Google" and "Hey Google" datasets, selecting a subset of 665 people, and verified it with a simple neural network which consisted of 1 linear layer and 3 LSTM layers. The results of the study were measured in Equal Error Rate (EER); researchers managed to reduce this indicator from 1.72% to 1.48%.

Researchers	**Scientific Approach**	**Achieved TAR**
Nematollahi, M. A. et al.[37]	MFA model developed based on online speaker recognition and multipurpose speech watermarking technology	**99.5%**
Adam Dustor [14]	GMM-UBM with VAD preprocessing and MFCC feature extraction	99.27%
Musab T. S. Al-Kaltakchi et al. [4]	Exploitation of an I-vector with three fusion methods	96.67%
Zhang et al. [61]	GMM with VAD preprocessing and MFCC feature extraction	92.75%
F. Thullier et al. [52]	Naive Bayes classifier with LPCCs	83%

Table 2.2: Achieved TAR on the TIMIT Dataset

Table 2.2 represents the achieved TAR for TIMIT dataset from the mentioned researchers. The highest TAR result was 99.5%, presented from work [37].

2.1.3 Other Methods

Physical Biometrics

Other physical biometrics include eye retina, iris, fingerprint, palm print and hand geometry [27]. Fingerprint, eye retina and iris are considered as the most convenient

methods due to high availability of the appropriate sensors.

Fingerprints have historically been used as a unique feature of the human body; even identical twins do not share same fingerprint pattern [27]. As one of the most prominent implemented biometrics, fingerprint models continue to develop in different dimensions. Recent studies on fingerprint biometrics introduced new models using 3D features of the finger [35], testing the efficiency of the hybrid models with traditional password and PIN-codes entry [45] as well as infused systems with probability-based methods and behavioural biometrics [19].

Eye-based recognition requires different types of cameras: for iris based methods a simple RGB camera is enough, while specific equipment including microscope components are needed to obtain the retina images. Contemporary studies for eye biometrics focused on improving feature extraction and matching methods. The majority of works upgraded the recognition system using Haralick textural features and statistical methods such as Co-Occurrence matrices [53], Local Luminal Variance [21] and Generalized Discriminant Analysis (GDA) [50].

Behavioural Biometrics

Behavioural biometric authentication system rely on continuous time-dependent data. The advantage of behavioural models is that data can be collected from normal usage of common devices such as mobile phones, tablets and computers.

Mobile sensors such as a gyroscope, magnetometer and accelerometer produce and collect motion data of the user. Scientific works used these sensors to discover behavioural patterns of users and create recognition models. Such biometrical systems implement several machine learning methods such as SVM, K-Nearest Neighbors (KNN) and Decision Tree (DT) [15] and deep learning networks "Deepauth" with LSTM component [6]. Several approaches examine data coming from recently introduced smart technologies such as Virtual Reality (VR) Oculars. The study conducted by Zhang et al. [62] proposed a prospective authentication model with data coming from VR headsets, and achieved decent EER results close to 9.7%.

In the last decade of the 21^{st} century interest in behavioral biometrics increased

rapidly, due to developments such as Keystroke Dynamics [34], [24], which demonstrated that the duration and latency of keystrokes as users entered their IDs and passwords could be used as a reliable method of authentication. Recent works on Keystroke Dynamics mostly prioritize the detection of the most precise approaches and the implementation of the system in the real world. Researchers [8] and [48] provided a detailed analysis on the optimal metric and machine learning algorithm for achieving competent results. The study by Kalita et al. [25] suggested a prospective mobile phone based system on keystroke dynamics for accessing physical facilities. The achieved EER scores for three different datasets were 2.34%, 3.63% and 9.23% and the work [25] outperformed the prior state-of-art approaches in the first two datasets.

2.2 Multimodal Systems

All multimodal biometric authentication systems are based on the merging of several unimodal authentication systems and implement fusion techniques on the feature level, matching level and decisions level to generate a total authentication score. The score is compared with the threshold provided by the system to generate a determination of the authentication outcome.

Many studies have combined face and voice characteristics to create a multimodal authentication system to overcome the limitations of unimodal biometric systems. A study conducted by Abozaid et al. [2] used the common PCA approach for face detection and used Artificial Neural Network (ANN), SVM and GMM for classification part. The EER of the model was 0.62% in the case of combining two biometric models. While some researchers [2], [59] used personally collected static image and voices of users, the paper presented by Chowdhury et al. [13] used footage of external CCTV camera to create a biometric recognition system. Their multimodal system was based on Disentangled Representation learning-Generative Adversarial Network (DR-GAN) for face recognition and 1-D CNN with MFCC and Linear Predictive Coding (LPC) feature extraction method for voice authentication.

Other listed physical and behavioural methods are also combined for the various multilevel systems. Behavioural unimodal components are used to strengthen the systems by pairing with a physical component; for mobile phones the behavioural biometrics such as gait and keystroke dynamics can be mixed for authentication [54]. Physical components rely on the camera presence and the associated sensors commonly in use in real-world applications. Models based on physical biometrics can use specific features of the user, such as iris and retina [44], which can be captured by a single specific camera.

The majority of multimodal biometric systems consist of two modules and a fusion score calculated on the feature level. In the study conducted by Supreetha et al. [51] the multimodal biometric system implemented and tested the performance of all fusion levels distinctively on face and palmprint biometrics. Score-level and feature-level fusions outperformed sensor and decision levels and achieved accuracy higher than 92%. Another work by Bayram and Bolat [9] introduced four-component multimodal biometric authentication using face, voice, thermal images and human ear image. The model used multilayer perceptron (MLP), DT, SVM and Probabilistic Neural Network (PNN) methods for authentication process and the highest performance was obtained by SVM, with 98.65% accuracy.

All of the mentioned works present a high level of user authentication accuracy, which can be useful for an intruder if he wants to log in with an existing user of the system, since the proposed works do not take into account an attempt to forge data. In addition, the authors of the works describe not only the success of their methods but also possible vulnerabilities, like spoofing, bypassing and overloading, that should be eliminated in future works. To bring novelty to previous research and improve system security, the discussed multimodal authentication models can be improved by using more difficult-to-counter human biometric features, one of which is a facial thermogram. Since a facial thermogram is unique for each individual, when combined with the user's voice data, this authentication method is supposed to be more reliable, because the chances of forging a thermogram or stealing it are very minimal.

2.2.1 Replicated Paper

As the base for our multimodal biometric authentication system, the work provided by Zhang et al. [59] was chosen for the practical part of the thesis, due to the claimed high rates of recognition, and the clarity of the architecture and model components. The Zhang system was implemented for Android-based application using the Java programming language and composed by a framework consisting of two components: modules for registration and authentication.

The system utilizes the sensors of the mobile device running the application. The sensors capture the face and voice of the user, and compare it with the registration data in the database. Researchers [59] built separate models for face and voice recognition, and then combined them using a matching scores fusion. Unlike the previous approaches, other feature extraction methods for data were used here, namely MFCC for sound and LBP for the face. The constructed models were tested on data from several datasets, and compared with the results of other models [57], [7] that were reproduced and tested with the same data. At the end of the study, the results of this work excelled in terms of TAR, FRR, FAR indicators, and the code execution time was significantly less than in the previous works. Regarding the advantages of the multimodal system, the authors of the work showed that their approach outperformed unimodal systems by using a combined model of face and voice features recognition.

Database	**Model**	**TAR**	**FAR**	**FRR**
XJTU database	replicated paper	100%	0%	0%
	study [7]	99.61%	0.20%	0.39%
	study [57]	100%	0%	0%
TIMIT and Georgia Tech database	replicated paper	90.28%	9.29%	9.72%
	study [7]	89.14%	10.14%	10.86%
	study [57]	90.43%	9.14%	9.57%

Table 2.3: Achieved results on the different datasets with several models

Table 2.3 represents the results of different models for the collected by Zhang et al. [59] XJTU database and compared models performances on the TIMIT and Georgia

Tech Faces databases.

2.3 Spoofing and Optimization

2.3.1 Spoofing

All of the reviewed multimodal biometric systems are based on unimodal structures which then generate an averaging acceptance score. However, the main disadvantage of the listed methods is that researchers used data collected in experimental setup environments, in controlled settings, using high-resolution sensors. Real-world performance is unlikely to match the laboratory achievements, due to the observation that use of the application will take place in less-controlled settings, at variable distances, subject to greater levels of noise, using lower-resolution sensors. The circumstances of actual use would likely lower the classification performance of the system. It is further noted that in the published works, the performance of the suggested systems were not examined for resilience under possible malicious attacks.

Rodrigues et al. [42] conducted one of the first studies toward spoofing techniques of multimodal biometric authentication systems. Researchers configured a face and fingerprint based model and used three fusion schemes for scores calculation including Weighted Sum, Likelihood Ratio (LLR) and Bayes LLR in four sessions: no spoofing, spoofing fingers only, spoofing face only and spoofing of both. The FAR obtained for the spoofing of both biometrics exceeds 91% for all fusion schemes while the session with spoofing only the face feature resulted in the FAR higher than 88%.

Proposed imposter attacks by early works were considered as simple and unrealistic, such as stealing biometric data by faking sensors, using previous accessed user data, and re-usage of the image and voice recordings of a user [43]. However, contemporary researchers' aims are detection of different imposter attacks [33] or detection of anomalies [10] on the authentication system. A few studies are provided with some guidelines for the distinctive realistic types of attacks.

In the research conducted by Scherhag et al. [46], morphing techniques were used

as a spoofing attack on the face authentication system. The work was focused on metrics for the detection of possible imposter attacks on the system. With provided guidelines for the morphing techniques and comparison between manual and automated morph, the authors reached 80% similarity score of manual morph with original users. Ergunay et al. [16] conducted research toward the detection of spoofing attacks on the voice systems and proposed three methods for imposter voice attacks such as changing the frequency of the authorized user voice, synthesizing voices, and voice conversion. The synthesizing voices technique uses the Hidden Markov Model (HMM) which was proven to change the accent, tone and other characteristics of voice. Voice conversion was implemented by laptop function to convert male voices to female voices. The FAR indicator for the scenario with speech synthesis via logical access surpassed 96% for males and 81% for females.

2.3.2 Optimization

Since security is an important factor for the biometric authentication system, contemporary studies started to invest in prevention techniques. Anti-spoofing techniques can be implemented in different segments of the model (Fig. 2-2). The liveliness detection anti-spoofing feature is implemented at the sensor and feature levels and is still in the development stage for each of the unimodal components.

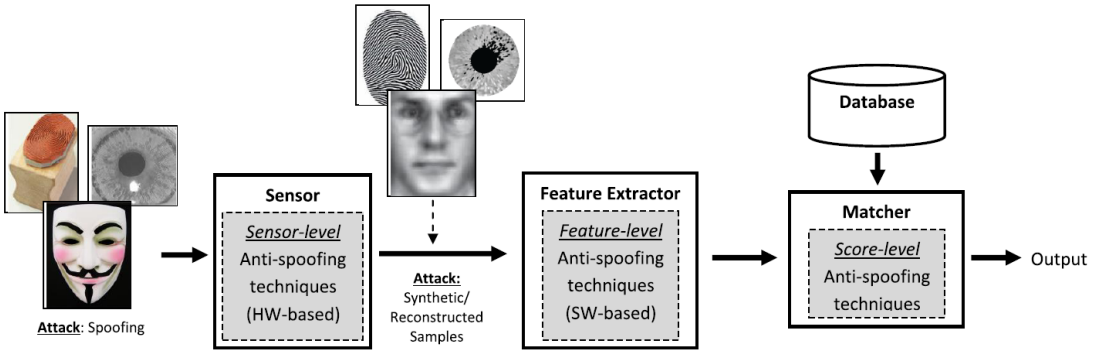


Figure 2-2: Anti-spoofing methods for different levels [17]

For behavioural components such as voice, keystroke dynamics and motions the

change in time-difference-of-arrival [58] can be computed, and anomalies detected. However, such systems remain susceptible to physical environment issues, and variations in the state of the user.

The liveliness detection component for the physical data sometimes required real-time authentication with 3D features. Weitzner et al. [56] completed a project on face recognition using a prototype of authentication system that distinguishes a real user's face from the image on different media (printed copy, phone) using 3D images or recurrent capture of 2D images for a particular time period. The disadvantage of this proposed method is that an imposter can create a 3D model of the user's image and will be able to gain access to the system. Other studies focused on the improvement of the security level by modifying and adding new components in the model parameters on the feature level without collecting additional types of data [28].

In addition, researchers work on the optimization of multimodal biometric authentication system not only on the security level, but also on the availability and integrity of the system. The majority of works are focused on enhancing the models to achieve higher accuracy by testing and comparing different techniques [49]. Moreover, the multimodal biometric systems are analyzed on the performance of the fusion score and time-cost complexity of the algorithms [59].

Chapter 3

Datasets

This chapter describes the datasets that are used to create, test, and validate our multimodal biometric authentication system, based on facial features and voice recognition. SpeakingFaces is the primary dataset that is used for creating the authentication system, and testing in terms of its security, integrity and availability. The TIMIT and Georgia Tech Faces datasets are used to validate the performance of the system, and compare results with published results.

3.1 SpeakingFaces Dataset

SpeakingFaces is a multimodal database consisting of aligned audio, video, and thermal data streams collected from 142 subjects, each of whom is recorded uttering approximately 100 commands. The commands are representative of typical human-computer interaction with digital assistants such as Siri or Alexa.

This publicly-available large-scale dataset [1] was gathered by researchers at the Institute of Smart Systems and Artifical Intelligence (ISSAI) of Nazarbayev University to encourage research in machine learning as applied to problem domains such as multimodal user authentication and speech recognition.

3.1.1 Data Collection

142 individuals participated in the collection process of SpeakingFaces, consisting of 74 male and 68 female subjects. Data was collected from each subject in two distinct sessions, to provide diversity in visual presentation and audio tone. Each session was divided into two sub-sessions. In the first sub-session the subject was asked to read aloud the phrases from one of the dual screens in three different sitting positions: looking to the front, left and right. The visual content was created by capturing the individual from nine different perspectives - each sitting position was recorded in line, at the bottom and from the top respectively to the face position.

The data collection station was equipped with a high-resolution thermal camera aligned with an RGB camera and an integrated dual microphone. Commands were presented to the subjects using two screens in order to limit head movement. The audio content was collected by asking participants to pronounce common commands for virtual assistants taken from the public Thingpedia and Siri datasets; the resulting recordings averaged approximately 4:30 minutes duration per participant.

The second sub-session was used to collect the visual content of the subject in a quiet, non-speaking mode, using a setup similar to that of the first sub-session. Data was collected from 9 different angles of the subject's face. The session was recorded in the quiet mode without using any audio tracks from the microphone.

Figure 3-1: Example of an RGB image from SpeakingFaces dataset

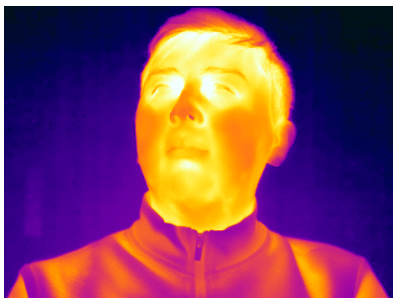


Figure 3-2: Example of a thermal image from SpeakingFaces dataset

The second session was conducted under the same instructions and setup as the first, but on a different day, with a gap of a week or more, at a different time of day, such that the subject was wearing different clothes, perhaps in a different mood, all of

which could affect the audio and thermal data streams. The variation was intentional, to introduce meaningful contrast to the data collection.

3.1.2 Initial Data Configuration

The visual stream of SpeakingFaces was collected as a standard RGB video file. Researchers wrote a script to divide each video file into frames. The video of each subject was split into approximately 900 frames for each angle. The scientists adjusted the RGB frames into the center in order to standardize the position of each subject for future works. The audio tracks were divided by individual spoken commands on a manual basis.

The dataset [1] is open-access, available by request from ISSAI; the full technical details can be found on their website. Data has been written into multiple folders by data categories that included video_only_raw, video_audio_raw, img_only, img_audio and uploaded into ISSAI server. The total data size is about 3.8 TB, which consists of 13,000 instances of spoken commands. All dataset files have their own code name consisting of several attribute values. It includes object id, trial id, session id, camera position, command number, frame number from video, image type and microphone number. For each individual participant, the data is stored in an individual .zip file.

3.1.3 Dataset Limitation

The SpeakingFaces dataset contains all of the media types of collected data for each object; however, some issues were detected in the data preparation stage, such as occasional missing files or malaligned images. Moreover, the variation in the angles of video collection introduced some irregularities, in that occasionally the facial images are cropped, or slightly blurred. In terms of audio, there is variation in sound quality between the two microphones, and for several subjects it is only the lesser microphone's data that is retained. The described omissions are present in a small part of the files, which does not greatly affect the quality of the dataset as a whole.

Thus, the SpeakingFaces dataset satisfies the objectives of this thesis and can be successfully used as a database of biometric characteristics of system users.

3.1.4 Dataset Preprocessing

Within the framework of the thesis work, we gathered a subset of the SpeakingFaces dataset and administered additional filtering procedures. The data from img_audio folder on the ISSAI server was downloaded for 111 users individually. The size of each zip folder of participant data ranged between 4.5 GB and 8.5 GB, with an average of 6.7 GB, divided into two session folders. Each session folder contained two audio folders from the first and second microphone and three visual content folder: rgb images, rgb aligned images and thermal images. As it was discussed in section 3.1.3, the final audio dataset was used only from the second microphone. In order to proceed with real-time data capture, for the purpose of our thesis we needed only the thermal and rgb images; the aligned rgb images were deleted as unnecessary for our purpose. Thermal and rgb images were manually extracted, retaining only one frame per angle for each user. The main criteria of the frame was full, non-cropped face with opened eyes and closed mouth. The whole process of downloading, cleaning and uploading a single participant's data took around 3-4 hours. However, by doing so, we were able to substantially reduce the volume of data needed, to an average of 37 MB per subject.

3.1.5 Subset Selection

For creation and validation of the performance of the multimodal biometric system, a subset of 111 users from the SpeakingFaces dataset was chosen. All users were selected randomly, since no description except the user id was given. The initially chosen 35 users had an imbalanced 1:6 gender ratio of women to men participants; the next 76 randomly selected users balanced out the gender ratio to 4:6. The average data size per user was less than 7.05 GB, as downloaded and preprocessed. Within the subset of 111 users, 13 users had minor issues with the visual data: the data collected

from some angles was fully blurred, such that each of those 13 had fewer than the 18 total distinct angles of images required for user authentication (9 vantage points from each of two sessions) for database creation and validation operations. Due to the insufficient amount of the image data, we determined to use these particular 13 users as the intruders group for analysis of spoofing techniques on the authentication system.

3.2 TIMIT Dataset

The TIMIT dataset is used to validate the audio component of the multimodal authentication biometric system. This dataset is one of the most popular datasets used in audio unimodal biometric systems.

The TIMIT acoustical-phonetic corpus [18] was intended for broad phonetic research, as well as for the development and testing of automatic continuous speech recognition systems in the framework of the American version of the English language. Several well-known organizations and research centers took part in the collection and development of dataset, including the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), Texas Instruments (TI) and National Institute of Standards and Technology (NIST). This dataset is assumed to be the one of the first speech corpora to be distributed on CDs.

3.2.1 Data Collection

Overall 630 speakers from 8 regional dialect zones of the USA took part in the recording of the corpus, where the developers strove for the same percentage distribution of dialects, although this criterion was not satisfied for all zones. The gender ratio of the speakers was also maintained, having each recorded dialect being represented by about 70% of male speakers and 30% of women. Among other features while selecting and recording speakers, researchers took into account age, height, race, educational level and speech recording time.

Pre-assembled text material was used to collect audio records of the dataset. The

TIMIT corpus text material includes 2,342 individual sentences. Among them, two sentences are specially constructed using phrases saturated with contexts in which one can expect the maximum manifestation of the speaker's dialectal affiliation. The remaining 2340 sentences are divided into two groups as follows. The first group consists of 450 special phonetically balanced sentences that provide full coverage of the phonemic inventory and the occurrence of phonemes in special contexts. For the second group 1890 sentences were selected from the available text corpora with the selection criterion to increase the variety of sentence types and phonetic contexts of the use of phonemes. The corpus uses 61 phonemes that are considered highly detailed and, for historical reasons, have been mapped into 48 phonemes for learning and 39 phonemes for testing and comparison among the scientific community.

3.2.2 Initial Data Configuration

In the TIMIT corpus, audio files received from different speakers are divided into training and test parts. The training part consists of 3696 pronunciations of 462 speakers and a test set of 1344 pronunciations of 168 speakers. During division of data into train and test sets, dataset developers guided special considerations. As a result, none of the speakers participated in both parts at the same time, so each part contains representatives of all dialects of different genders. The training and test data contain unique sentences; the test data provide full coverage of the phonemic inventory, a sufficient variety of their phonetic contexts and frequency of occurrence. Each sentence, both in the training and test set, is associated with four different files, which differ only in extension and contain different information about the spoken sentence. One of the files is sound, and the rest are text. The structure of the associated text files is the same and reflects the time-alignment of different language objects with the signal, that is, different levels of its markup.

3.2.3 Data Preprocessing

For further recognition of the system user, all the data used were taken only from the TRAIN folder of the dataset, in the amount of 462 people, since the test sample does not contain separate records of each individual for testing the model. For each speaker, the audio files were divided in a ratio of 8:2 into train and test data, respectively, thereby preserving the presence of samples of all users in both cases.

3.3 Georgia Tech Face Database

The Georgia Tech Face Database [36] is the validation dataset of the face feature based unimodal biometric authentication system which is a part of the whole system. This dataset is still in use in the modern face-based biometric authentication systems.

The database contains the data of 50 different participants whose faces were captured in the period between June and mid-November in 1999. The data collection was conducted in the Center for Signal and Image Processing at Georgia Institute of Technology. The database size is approximately 128 MB, and is freely available on the internet. Each subjects data consists of 15 RGB images with furniture on the background: shelves, sculptures and other. All pictures were collected in the same size of 640*640 pixels.

3.3.1 Data Collection

Each individual that participated in the data collection process was asked to come for one to three sessions. However, despite the number of sessions for each individual, the database stores an equal number of samples for each user. The major differences in the images between sessions are scaling factor of the face and lighting conditions. In order to get the variety of data, each participant was captured with various facial expressions: smile, laugh, closed-eyes, etc., while the face position on the image remained in almost the same frontal angle. Moreover, researchers labeled manually every single image from 1 to 15.

Figure 3-3: Example of an image of the Georgia Tech Face Database

Figure 3-4: Example of similar facial expression image from Georgia Tech Face Database

3.3.2 Data Preprocessing and Limitations

For the validation purposes the Georgia Tech Face Database was manually split into training and test sets with ratio 80:20. The first 12 labeled images were used as a training set and 3 other images validated the peformance of the system.

The main limitation of Georgia Tech Face Database is the similarity of the expressions between the sessions. As a result, images with the same expressions were manually divided into test and train data as it can be seen in the Figure 3-3 and Figure 3-4. Since the data contained the frontal position of the face, only the frontal module of face feature recognition system was validated by using the Georgia Tech Face Database. Moreover, there is a lot of furniture included in the images and some additional regulations for the system were required to recognize the face of the participant.

Chapter 4

Methodology

As noted in chapter 2, the original architecture of the multimodal biometric authentication system was created by Zhang et al. [59]. This chapter provides a detailed overview of the original work model, the parameters and setups for our replication of that model, a description of the differences between the two, and a brief description of the model's limitations.

4.1 Original Model

The biometric authentication model proposed by Zhang et al. [59] was created on the Android platform using the Java programming language. Similar to other multimodal systems, the original architecture is comprised of two unimodal components which are the linked with a specific fusion score algorithm.

The architecture of the user authentication model presented in Fig. 4-1 consists of 2 separate face and voice biometrics-based unimodal classification models. The model was tested on the XJTU multimodal database of 102 individuals. In the process of face matching, after data preprocessing, the face area on the images is determined using the Haar + AdaBoost methods. The necessary features of a face are acquired using the Local Binary Patterns method and then used in the Euclidean metric to determine the similarity to an existing face in the dataset. The matching score is kept until the combining stage.

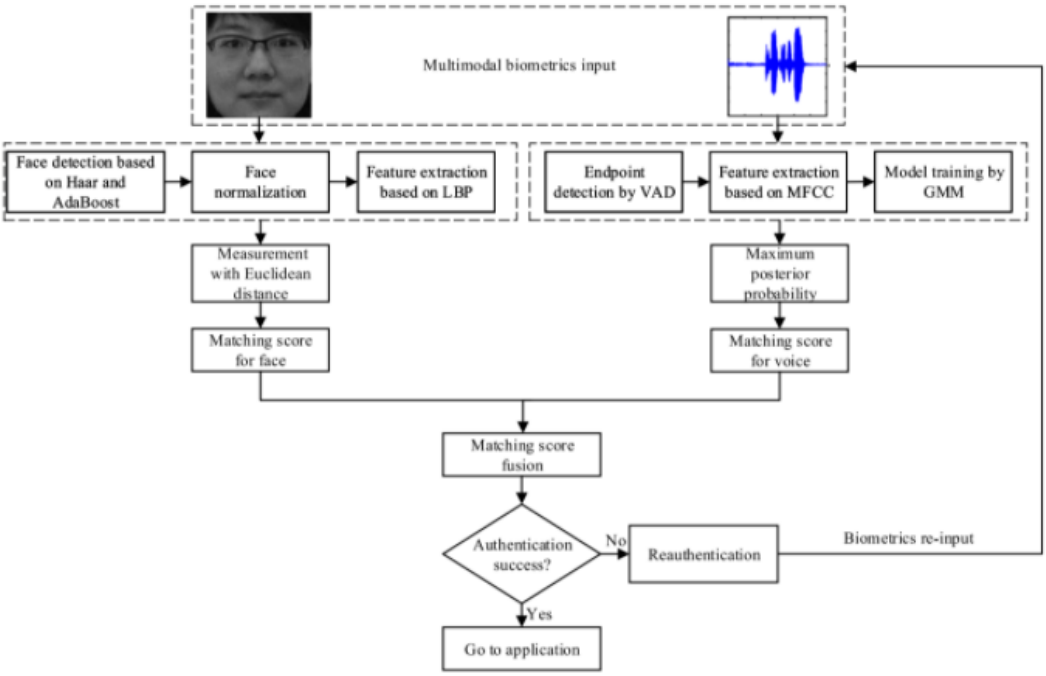


Figure 4-1: Flow chart of the multimodal authentication system implemented by Zhang et al. [59]

For voice recognition, data is preprocessed with Discrete Wavelet Transform for noise reduction. Useful components are isolated from sound files using the Voice Activity Detection method, where a person is talking without noise and hesitation, while silence is ignored. With the help of the Mel-Frequency Cepstral Coefficients, features are obtained from the cleaned data, and the classification is passed in the Gaussian Mixture model, where the matching score is also computed as with the face verification. The results are then matched with scores fusion to test both user characteristics. The parameters TAR, FRR, FAR and program running time are used to evaluate the results.

4.2 Model Components

4.2.1 Face Module *(AI)*

The facial features recognition module for the Android based multimodal biometric authentication system [59] consisted of two procedures: the creation of the training database and the testing of the recognition system. From Fig. 4-2 , it can be observed that creation of the training part divides into three stages: image preprocessing, face detection and facial features extraction. The image preprocessing stage was not described in full; however, the prior article of Zhang et al. [60] provides a simple description of the preprocessing stage: histogram normalization, medium filtering with kernel size = 3 and pixel value normalization.

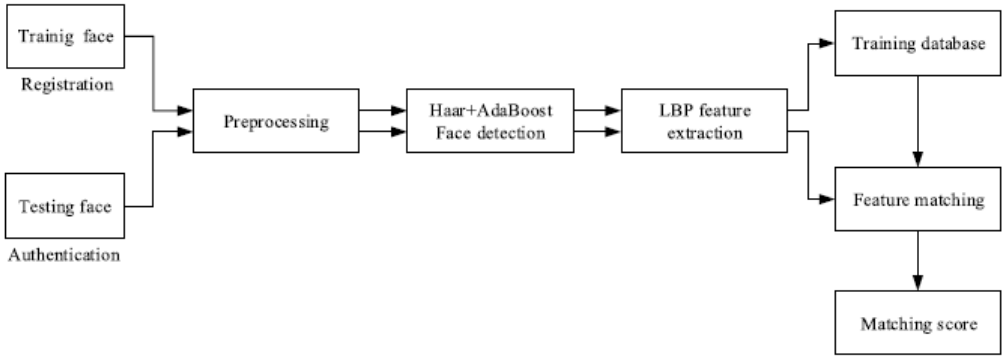


Figure 4-2: Unimodal biometric component for face recognition by Zhang et al. [59]

Haar Cascades

The face detection process consists of the combined Haar Cascades algorithm of Haar features computation and AdaBoost training. The Haar features calculation is essentially a computation of the pel intensity magnitude within specific image area. The area is represented by the adjacent rectangular shaped regions with a particular position of the sliding window in the region. The computation includes the sum of pixel intensities at specific regions, and difference of the sums.

However, the computation of Haar features for large images was resource-intensive in terms of both time and memory. The introduction of the integral division of the

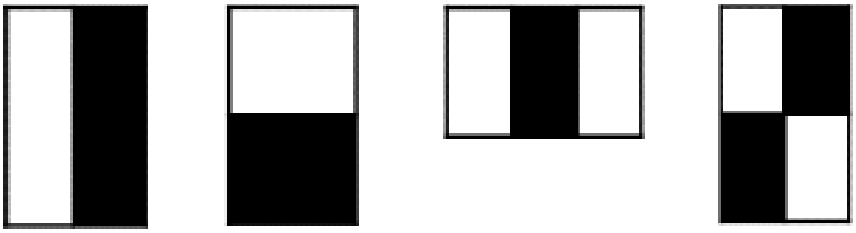


Figure 4-3: Haar features templates.

image sped up the computation of the Haar features by creating rectangular sub-regions and array references for each sub-region. Due to integral division of the image, the recent calculation of Haar features uses rectangular regions instead of pixels for the large pictures.

After merging the notion of integral images and Haar cascades the issue of the improving object detection was considered. The object detection procedure required highlighting regions of importance. The solution for improving the detection methods was to utilize the Adaptive Boosting (AdaBoost) technique. AdaBoost is well-known algorithm for boosting binary decision trees; nevertheless, this algorithm is recently used for improving the performance of many machine learning algorithms. The Fig.4-4 represents the main concept of the AdaBoost classifier: it increases the accuracy of the classification problem by creating a strong classifier from a number of weak classifiers.

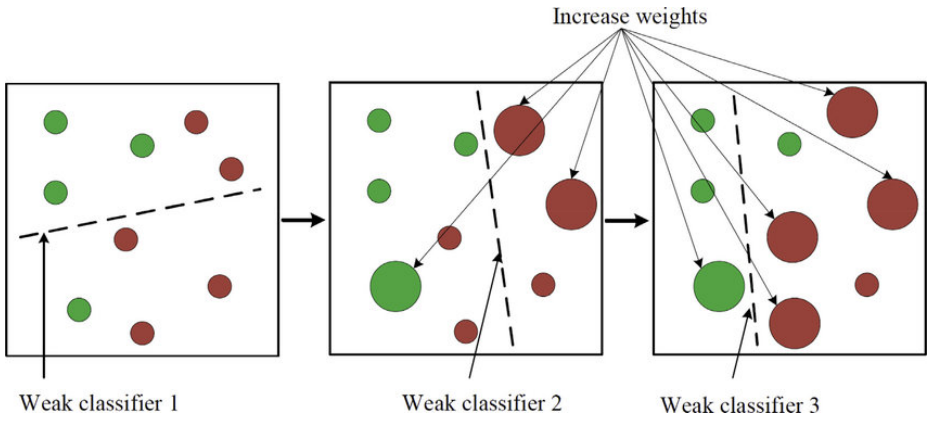


Figure 4-4: Simple principle of AdaBoost work

The adoption of the AdaBoost classifier for object detection with Haar integral features resulted in the creation of the famous Haar Cascades method for object detection. The architecture of Cascade Classifiers on the Fig 4-5 represents the process of the classifier, where at each stage the collection of weak classifiers determines the existence of the object. In the case when an object is not detected, the process moves to another stage by using different weights of the weak classifiers until the object is detected.

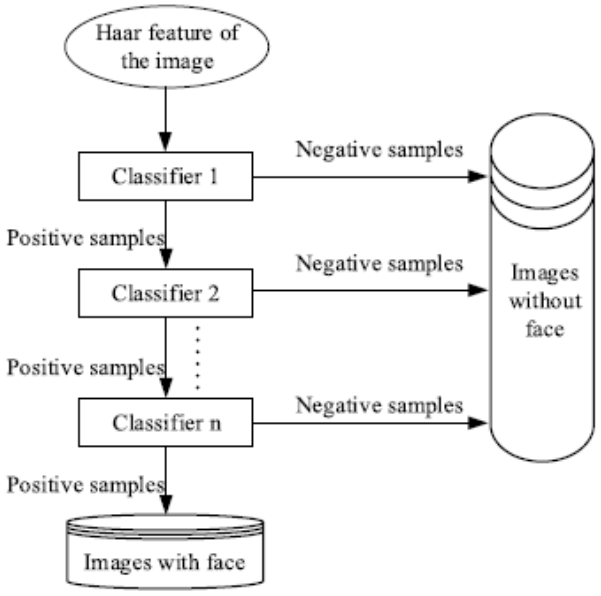


Figure 4-5: Representation of Haar Cascades Algorithm

LBP Feature Extraction

Following the face detection phase, feature extraction is implemented to minimize the impact of noise and reduce information redundancy by saving only important features. The Local Binary Pattern (LBP) texture operator is a powerful and straightforward for implementation algorithm which recomputes the value of each pixel by implementing a threshold for neighborhood pixels. Due to its discriminative power, perfect representation of the features and computational integrity, this operator has become a favored approach for variety of machine learning methods.

Before applying the LBP coding principle, the image is initially converted to the

gray-scale and then the value of each pixel in the neighborhood is recalculated based on a comparison with the central pixel. The formula of LBP coding principle can be represented as:

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} 2^p * s(i_p - i_c), \tag{4.1}$$

where $s(x)$ is the sign function for x$\geq$0 then $s(x)$=1; otherwise, $s(x)$ =0. The P represents sets of the neighbor points within radius R and p is a particular point from the set P.

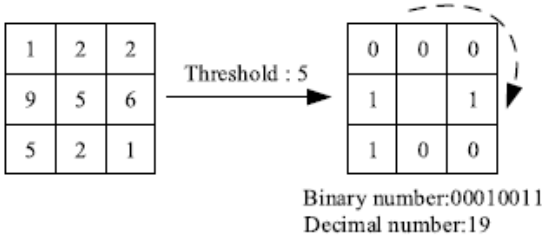


Figure 4-6: Example of LBP coding

Zhang et al. [59] used the improved version of the LBP coding algorithm on Eq.4.1 which reduces category number of LBP 2^p part with $p(p-1)+2$. The improved version of algorithm was shown to provide the invariance in the grayscale as well as in the rotation of the feature image. The whole feature extraction algorithm proposed by the authors [59] can be described as follows:

- 1. Apply LBP operator and compute feature image.
- 2. Division of the feature images into uniform blocks
- 3. Computation of the LBP code histogram for each of the block from Step 2.
- 4. Create the feature vector by concatenating the histogram features of all blocks according to the spatial order.

4.2.2 Voice Module *(AK)*

The unimodal voice identification component is similar to the unimodal face component in that it is combined from two processes. Fig. 4-7 represents the stages of the voice matching process (described in more detail below): noise reduction using DWT, preprocessing of voice data, silence detection based on the VAD method, extraction of the voice features using MFCC and creation of the voice identification model based on GMM.

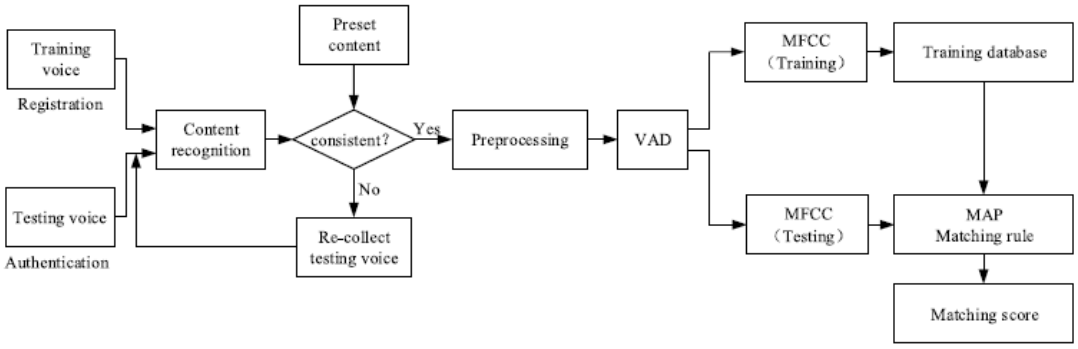


Figure 4-7: Unimodal biometric component for voice identification by Zhang et al. [59]

Discrete Wavelet Transformation

Discrete Wavelet Transformation (DWT) is widely used in the analysis and synthesis of various signals. With DWT the original signal is passed through a high-pass filter and a low-pass filter. The output of the filters will be respectively high-frequency (HF) and low-frequency (LF) signal components. For many signals, the low-frequency component is the most important from the point of view of the information it conveys. For the most part, it repeats the signal itself, but is not an exact copy of it. Therefore, the low frequency coefficients are commonly referred to as "approximation" coefficients. The high-frequency component, on the contrary, is less informative, and its task is to supplement the low-frequency signal. It is also commonly called the "detailing" component.

Processing and cleaning the audio signal from noise occurs in several stages:

- Decomposition. The wavelet and the decomposition level are selected, and the wavelet decomposition of the original signal is calculated.
- Detailing. The threshold is determined and the resulting detailing coefficients are processed.
- Reconstruction. A wavelet reconstruction is performed based on the original approximating and modified detailing coefficients.

At the first stage of conversion, a digital filter separates the signal into low frequencies and extracts high frequencies. Since there is no upper half of the frequencies at the output of the low-pass filter, the sampling frequency of the output signal can be reduced, i.e. the procedure of "decimation" of the output signal has been performed. At the output of the high-pass filter, space is freed up in the low frequency region, and a similar decimation of the output signal leads to the transposition of the high frequencies to the vacant space. This DWT algorithm for audio file denoising was implemented with PyWavelets, a free open-source library for wavelet transforms in Python.

Voice Activity Detection

Voice Activity Detector (VAD) is a method for determining the activity of speech, which is a technology for compressing a speech signal by searching for speech and pauses segments, and encoding them. The performance and efficiency of the authentication system with speech signal recognition can be primarily examined with the effectiveness of the implementation of VAD algorithm. The presence of pauses is determined based on the analysis and synthesis of speech data that contain signal segments.

In the most common implementation of the algorithm, during comparison of the signal frequencies of the initial data signal, the threshold value which divides the frame with silence and with actual voice in the audio signal into separate parts is used. In this case, the threshold is selected in such a way as to prevent excessively frequent elimination of erroneous pauses, as this can lead to a deterioration in quality,

loss of important data and, as a consequence, to a decrease in the efficiency of the VAD algorithm. In more advanced implementations of VAD, a complex algorithm is used to determine the pauses, where the spectral components of the signal are used along with its energy frequencies.

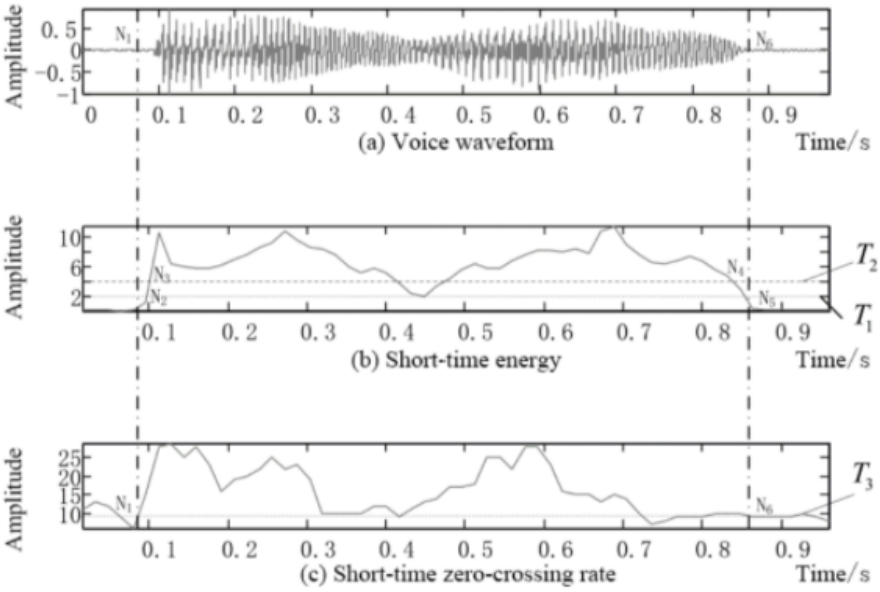


Figure 4-8: Principle of the improved VAD method by Zhang et al. [59]

The authors proposed improvements to the algorithm reflected on Fig. 4-8 due to the weak efficiency of the method for signals where the peak of the sound amplitude in relation to the noise level is very small [59]. The idea behind the improved VAD algorithm is as follows. Initially, two energy thresholds are selected: high - T1 based on the average signal energy and low - T2 determined using the background noise level. Further, the obtained values are used to roughly determine the start and endpoints of the audio signal. Finally, the zero-crossing threshold T3 is calculated based on the average zero-crossing rate of the noise part, which allows for the recognition of more accurate speech endpoints.

Mel-Frequency Cepstral Coefficients

This work applies the most popular feature extraction method by calculating Mel-Frequency Cepstral Coefficients (MFCC). The cepstral transformation coefficients

form the space for speech recognition. The scheme of this method is as follows: on a time interval of 10 - 20 ms, the current power spectrum is calculated with next application of the inverse Fourier transformation of the logarithm of this spectrum (cepstrum), and at the end the cepstrum coefficients are found. The number of cepstral coefficients depends on the required spectrum smoothing.

Finding MFCC is carried out in several stages:

1. Normalization of the original signal to equalize its amplitude and increase of high frequencies.

2. Isolation of a short-term signal section (frame) and overlay of a window function to minimize spectrum leakage.

3. Computing the DFT of a frame for which the Fast Fourier Transform (FFT) algorithm is used.

4. Application of a set of mel-filters on a frame, where speech perception is simulated similar to a human, in a way that the hearing resolution grows when moving along the spectrum from low frequencies to high

Hearing properties are taken into account by non-linear transformation of the frequency scale called the mel scale. This scale is formed based on the presence of so-called critical bands in the ear, such that signals of any frequency within the critical band are indistinguishable. The frequency transformed into mel scale is calculated as:

$$f_{Mel} = 2595 \log(1 + f/700)$$

where f is the frequency in Hz, f_{Mel} is the frequency in mels.

Gaussian Mixture Model

Due to the fact that the vast majority of speaker recognition systems use the same feature space in the idea of cepstral coefficients of the first and second differences, the main attention is paid to the construction of decision rules, for which the method of

approximating the probability density in the feature space with weighted mixture of normal distributions (GMM).

The Gaussian Mixture model assumes that the data is subject to a Gaussian mixture distribution, in other words, the data can be viewed as generated from multiple Gaussian distributions. The application of the GMM method for voice recognition is often employed in context-independent systems for the description of the speech features' probability density and their approximation for the target speaker.

Each GMM consists of Gaussian distributions, and each Gaussian is called a component. Linear addition of these components forms the GMM probability density function:

$$p(x) = \sum_{k=1}^{K} c_k \varphi(x|\theta_k),$$

$$\sum_{k=1}^{K} c_k = 1$$

where $\varphi(x|\theta_k)$ is the distribution function of the sample x with parameters θ_k, the vectors of the weights c_k and the number of components k.

With this approach, the initial data are presented in the form of clusters described by Gaussians. Usually, a fixed number of mixture components is specified, and the components' principal axes are placed in the same direction with the feature space coordinate axes. This is due to a large number of computations, which are shortened using the diagonal covariance matrix.

After finding the closest model, it is necessary to refer the audio sample to a registered or unregistered user. For this the method of maximum posterior probability is used to evaluate the parameters of the mixture. Based on the comparison of this value with the threshold, a decision is made about the speaker to be tested.

$$P(\theta|x_0) = \frac{P(x_0|\theta)P(\theta)}{P(x_0)}$$

4.2.3 Fusion Score

The grouping of unimodal components into the multimodal authentication system can be processed on different levels such as the feature level, matching level and decisions level. Due to the significant differences between voice and facial features, the original paper chose the matching level for the composing the fusion score and introduced normalized f_{score} and v_{score} which stands for normalized scores using min-max approach of facial and vocal features respectively. Figure 4-9 represents the flowchart of the fusion calculation process in the multimodal biometric authentication system, where distance score corresponds to f_{score} and likelihood score stands for v_{score}.

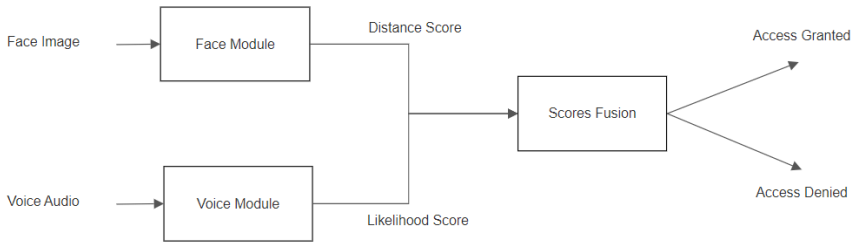


Figure 4-9: Flowchart of the fusion score calculation process

The fusion score calculation is based on two fusion tasks t_1 and t_2 which are computed as:

$$t_1 = a * v\_score + (1 - a) * f\_score,$$
$$t_2 = (a * v\_score) * ((1 - a) * f\_score),$$

where a is the weighted value. The value of a is directly dependent on the data quality: for noise-free audio and visual data the value of a is equalized to 0.5 in order to not discriminate each unimodal component. The final decision fusion score is evaluated as:

$$f_{decision} = f(t_1) * f(t_2)$$

where $f(t_1)$ and $f(t_2)$ are step functions of the sum and product decisions respectively.

$$f(t_1) = \begin{cases} 1, t_1 \geq t_{sum} \\ 0, t_1 \leq t_{sum} \end{cases}$$

$$f(t_2) = \begin{cases} 1, t_2 \geq t_{pro} \\ 0, t_2 \leq t_{pro} \end{cases}$$

t_{pro} and t_{sum} represent the threshold values for the production and sum fusion scores respectively. The final decision of authentication system depends on the $f_{decision}$ value.

4.3 Limitations of Prior Work

Certain gaps and limitations were observed in the original work [59], such as the absence of the parameters for each unimodal component as well as not covering the threshold identification process.

4.3.1 Visual Part *(AI)*

On the part of the face recognition module, several weaknesses were identified during the system testing phase on the SpeakingFaces dataset and Georgia Tech Face Database.

The limited description of the image preprocessing stage was one of the main obstacles to reproducing the biometric authentication system by Zhang et al [59]. Despite the fact that the SpeakingFaces dataset has minimal redundancies and noise in the visual data, the face detection method was problematic due to profile face shots of the user. The existing profile Haar features template still needs improvement and can be only applied for the left profile of the face. Additional procedures of double mirroring the right profile of the user were added before and after the face detection stage. However, the prototype of the frontal Haar template has issues with correctly defining the face area of the participant.

The most significant issue was discovered during the face matching process of the

authentication to the system. The face region of each participant in the SpeakingFaces training subset was different in size after the face detection phase of Haar cascades. The work by Zhang et al. [59] did not mention whether the processes of reshaping feature images or whitening the non-facial background of the images in order to keep same image size were used. Nevertheless, for the recreated model the reshaping feature image in the database was implemented in the face matching process during testing of the database.

4.3.2 Audio Part *(AK)*

On the part of the voice recognition module, potential weaknesses of the applied methods related to the quality of the tested SpeakingFaces dataset were also discovered.

During preprocessing an audio signal with DWT, each of the resulting signals carries information about its part of the frequencies, while the output information is represented by the same number of samples as the input, sufficient to reconstruct the signal. Due to the high noise level during the recording of audio files of some participants, one-level conversion is not enough to remove noise from the signal or implement its compression due to the loss of a large number of useful components of the original signal. The potential solution is to resort to the procedure of repeated wavelet decomposition.

The necessary characteristics for an ideal voice activity detector are reliability, stability, accuracy, adaptability, simplicity and ability to use without information about the noise present. During audio data collection, resistance to noise is the hardest criterion to achieve. Under conditions of high signal-to-noise ratio (SNR) of audio files in the Speaking Faces dataset, the simplest VAD algorithms work satisfactorily, but under low SNR conditions, VAD algorithm degrades to a certain extent. At the same time, the VAD algorithm must remain simple to meet the requirement for real-time applicability. Thus, the input signal must be carefully cleaned of noise so as not to create problems during the segmentation of useful frames.

4.3.3 Fusion Score

As it was mentioned before, the major issue in replication of the prior work was the absence of required parameters. The fusion score as **introduced by Zhang et al** is heavily dependent on the thresholds for each parameter. Since all values were normalized for the original work [59], it was difficult to recreate the same fusion score scheme without knowing the thresholds. Therefore, the replicated model was tested on two fusion scores: threshold based with own defined thresholds and ranking based fusion score which calculates the rank of similarity of each unimodal stage and provides access to the user according to whether he/she passes the rank test. The Ranking based fusion score was implemented as an alternative fusion score for comparison purposes to the threshold based fusion score results. Both vocal and visual features of a person are compared to all data in the training databases and each training data creates a ranking of the similarity for each user. For the purposes of this work only the top 5, top 7 and top 10 closest users were chosen for creating the new fusion score system. The rationale for selecting these particular ranking numbers is the potential vulnerability of the system: as more similar users are allowed to enter to the database it increases the possibility of successful intruder attack. Smaller numbers were not considered due to low performance received in the testing stage.

4.4 Parameters and Initial Setup

4.4.1 Facial Module *(AI)*

In the facial module, each component of the authentication process requires its own specific parameter values. The image filtering stage of the original model proved unnecessary for the SpeakingFaces dataset, due to the specificity of the data collection. All users were filmed against a white background with negligible noise and redundancies. For the face detection phase with usage of Haar Cascades based in AdaBoost algorithm there are four common templates that were available for the Haar features detection: default, alternative, alternative 2 and tree of alternative templates. More-

over, for the AdaBoost algorithm parameters such as the scaling factor and number of neighbor features need to be defined for correct object recognition. The extraction process of the face features is used for the improved LBP coding system with recalculating the neighbor pixels around. Since the sizes of the face detected on the training set vary from user to user, a specific reshaping module was used for the testing module. Two setups were generated for the SpeakingFaces dataset: with using histogram computation and without.

The values and ranges of the parameters of the visual module are set as follows:

- Haar templates: default and alternative2.
- AdaBoost parameters: scaling factor in range within 1.001 and 1.01 with number of neighbors in range between 3 and 34
- LBP: number of neighbors is 8 within radius R=1
- Reshaped size of the face image: $150 * 150$ pixels

4.4.2 Speech Module *(AK)*

In the speech module, each stage of building the authentication system uses its own configurable parameters. At the stage of removing noise from audio, the value of the threshold is used as a tuning parameter, which is the border of separation of a high-quality signal from background noise along with the type of wavelet used for cleaning audio. Further, to recognize the voice activity on the audio signal, the upper amplitude, the frame length and the hop length are selected, where subsequently all recognized voice fragments are concatenated into one separate file. When extracting features of a speech signal, the signal itself and its rate, the length of the analysis window, the step between successive windows, the number of cepstral coefficients and the size of FFT are used as parameters. At the next stage of building the GMM classification model, the main tuning parameters are the number of mixture components, the type of covariance parameters, the number of EM iterations and the number of initializations to perform.

The range of parameters' values of the speech module are set as follows:

- DWT parameters: $threshold \in \{0.05, 0.15\}$, $wavelet =' sym4'$
- VAD parameters: $top\_db = 20$, $frame\_length = 2048$, $hop\_length = 2$
- MFCC parameters: $winlen = 0.025$, $winstep = 0.01$, $numcep = 20$, $nfft = 1200$
- GMM parameters: $n\_components \in \{50, 300\}$, $max\_iter \in \{50, 300\}$, $covariance\_type \in \{'tied', 'diag'\}$, $n\_init \in \{1, 3\}$

4.4.3 Fusion Score

The fusion score calculation is provided in two ways: ranking based and the original fusion score with thresholds. Our multimodal system obtains results on the voice and face biometrics from unimodal components and uses them in the calculation process. The ranking based system only requires the scores of similarity of the provided biometric credentials with other users in order to determine whether it is an authorized user and which data belongs to that user. The single parameter that is needed for the ranking based fusion score is the top rank of users that will be considered.

The prior fusion score took as input the likelihood scores of biometrics with users in the database and a username credential. Three parameters can be modified in order to test the performance of the system. The weighted value a is set as 0.5 for equal treating both unimodal components of the authentication system; however, the values of sum and product fusion scores will vary.

- Ranking: top5, top7, top10
- t_sum: in range within 0.01 and 0.1
- t_prod: in range within 0.01 and 0.1

Chapter 5

Results

This chapter presents the results achieved used the model architecture as described in the chapter 4 and in the Appendix A. Our multimodal biometric system was developed using a subset of SpeakingFaces, and validated using the data as described in the section on Datasets.

The unimodal components were developed and tested in the model across a range of system parameters. The parameters associated with the best outcomes of the unimodal systems were then transferred to the multimodal system, and the overall performance evaluated using the SpeakingFaces data.

5.1 Unimodal Components

5.1.1 Facial Recognition *(AI)*

For testing of the SpeakingFaces subset, different model setups were initially assessed using only frontal profiles of the participants. As it was discussed in Ch.4, there are four Haar cascades templates that can be used in the frontal face detection step; in our experience, the alternative2 (or alt2) template demonstrated the highest accuracy rate for face recognition - 100%. This template was in use for all parameter testing.

For evaluation of the effect of the number of neighbors used in face recognition on the accuracy of the model, several values were tested. Table 5.1 illustrates the

behaviour of the model's top 5 accuracy results on changes in the number of neighbors with scaling factor 1.01 and the Euclidean metric for distance calculation. The number of neighbors is used to improve the identification of the face within the image; increasing the number of neighbors improves the recognition of the facial region but may diminish the recognition rate.

It is clearly observed that there is no difference in results for a number of neighbors between 3 and 11, while there are slight improvements in all parameters for 18 neighbors rectangle calculations.

Number of neighbors	3	4	6	8	11	18
TOP1	69.05%	69.05%	69.05%	69.05%	69.05%	69.39%
TOP2	72.79%	72.79%	72.79%	72.79%	72.79%	73.13%
TOP3	76.53%	76.53%	76.53%	76.53%	76.53%	76.87%
TOP4	78.57%	78.57%	78.57%	78.57%	78.57%	78.91%
TOP5	79.59%	79.59%	79.59%	79.59%	79.59%	79.93%

Table 5.1: Achieved TOP5 accuracy for Speaking Faces for various number of neighbors with scaling 1.01 and Euclidean distance

Table 5.2 represents the dynamics in the performance of the model with various scaling factors in the face recognition step. The system was evaluated with Euclidean metric distance and best neighbor numbers - 18. Overall, the reduction in the scaling factor led to an insignificant reduction of the top 5 accuracy results on the frontal faces of participants.

Scaling Factor	1.01	1.005	1.001
TOP1	69.39%	69.39%	68.71%
TOP2	73.13%	73.47%	72.45%
TOP3	76.87%	76.87%	76.87%
TOP4	78.91%	78.23%	78.57%
TOP5	79.93%	79.25%	79.25%

Table 5.2: Achieved TOP5 accuracy for Speaking Faces for different scaling factor with 18 neighbors and Euclidean distance

The original model by Zhang et al. [59] recommended and validated all results on the Euclidean distance metric. However, the recreation of their model used with the Speaking Faces dataset compared the performances of the Euclidean distance

and City-block or Manhattan distance. Table 5.3 demonstrates the difference in the achieved accuracy between two metrics with established best parameters: 1.01 scaling factor and 18 neighbors; the City-block metric outperformed the Euclidean metric.

Metric	Euclidean	City-block
TOP1	69.39%	71.09%
TOP2	73.13%	74.49%
TOP3	76.87%	78.23%
TOP4	78.91%	80.61%
TOP5	79.93%	81.29%

Table 5.3: Achieved TOP5 accuracy for Speaking Faces for different metrics with 18 neighbors and 1.01 scaling factor

The major differences between the prior work [59] and our recreated model were based on observations of the data, and the considerations of the relative impact of each stage of the feature extraction process. The SpeakingFaces Dataset was collected in a laboratory environment, with little or no noise (refer to Ch.3), thus, the preprocessing stage used in the prior work was not required. However, after testing the preprocessing procedures of histogram normalization, medium filtering with kernel size = 3 and pixel value normalization on the SpeakingFaces frontal face subset, the recognition rate fell to 68%, failing to detect any faces on some sets of the images. The three stages of the image feature extraction process consist of dividing on the blocks, calculating histograms and creating the feature vectors; the latter two steps were excluded from the final model due to significant drop in the top 5 accuracies: TAR was 22.11% and top 5 was 45.24%. The final version of our system implemented only the distance calculation between features of the received image and that stored in the database buffer. Fig. 5-2 and Fig. 5-1 represent the feature image and incoming image for the model respectively.

After obtaining all required parameters for the model, the system was tested on the subset of the SpeakingFaces dataset; table 5.4 shows the results of the unimodal biometric face recognition system. As it was shown on the frontal faces subset, the City-block metric outperforms the Euclidean metric results. Nevertheless, the overall results of the whole trained dataset drops compared to the only frontal faces results

Figure 5-1: Example of an RGB image from SpeakingFaces dataset

Figure 5-2: Example of a feature image created after for database

having top 5 presence 74.15% and 81.29% respectively. A possible explanation for the diminishing performance is in the profile Haar Cascade template, which is presented in the single version and could be further improved.

Metric	**TOP1**	**TOP2**	**TOP3**	**TOP4**	**TOP5**
Euclidean	58.1%	63.76%	67.16%	69.99%	72.71%
City-block	60.77%	65.76%	70.18%	72.9%	74.15%

Table 5.4: Achieved TOP5 accuracy for whole subset Speaking Faces for two metrics

5.1.2 Voice Identification *(AK)*

At this stage, experiments were conducted on user audio record recognition quality by varying the parameters of the system components and observing the outcomes.

The main results were obtained from testing a variety of setups during the construction of a user recognition model. Table 5.5 shows the performance of the user recognition system built on the GMM model. The first column shows the names of the tuning parameters, in the following columns, tested combinations of their values are indicated. In the last two rows we show the accuracy of the user authentication and the probability of finding a designated subject among the top 5 most-similar.

From the constructed Table 5.5 observed an increase in accuracy values with an increase in the number of model components and a decrease in the number of iterations. So, for example, in Setup 1, the recognition accuracy is the lowest with the

Parameters	**Setup 1**	**Setup 2**	**Setup 3**	**Setup 4**	**Setup 5**	**Setup 6**
Number of components	50	100	150	100	300	150
Number of EM iterations	300	300	200	100	100	50
Type of covariance	'diag'	'diag'	'diag'	'tied'	'tied'	'tied'
Number of initializations	3	3	1	1	1	1
Presence within TOP 5	81%	87.7%	90.34%	91.08%	91.36%	92.18%
Accuracy	66%	70%	71.63%	73.4%	74.42%	77.55%

Table 5.5: Effect of GMM model parameters setup on the quality of user recognition

smallest number of components and the largest number of iterations. Talking about the type of covariance, the data fits better on the tied covariance matrix than on the diagonal. Overall, among all the tests carried out, the highest obtained recognition accuracy was 77.55% with the parameters' values used in Setup 6.

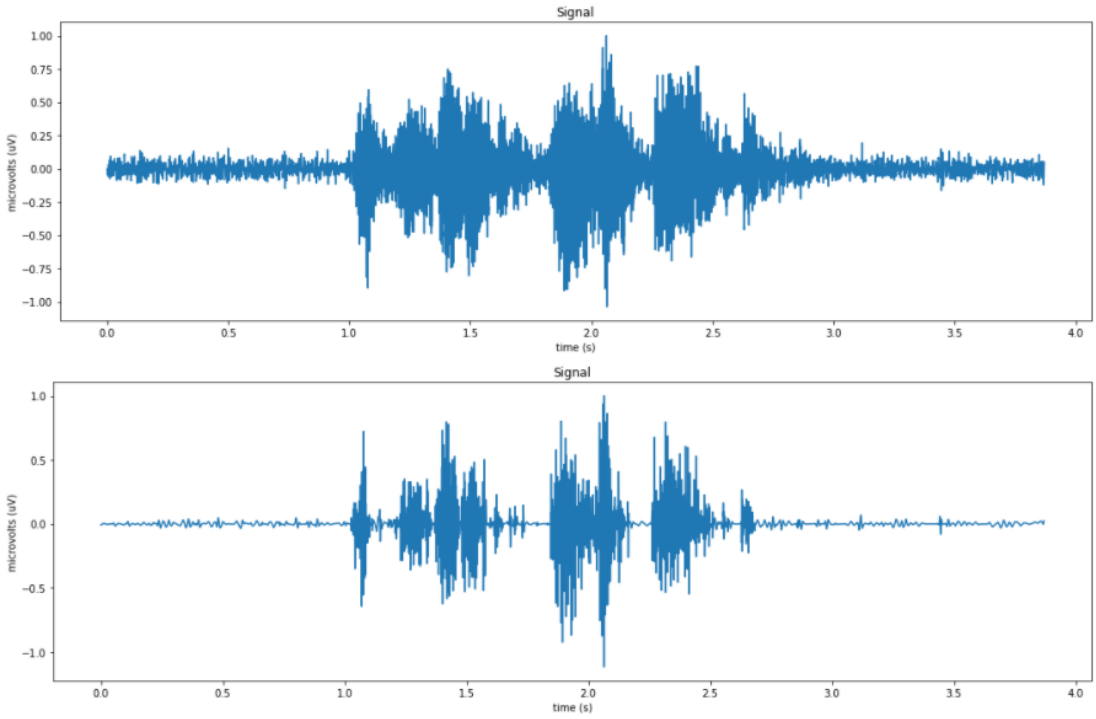


Figure 5-3: Example of DWT application on noisy signal

The results obtained during system testing were highly dependent on the presence of noise in the audio files. During testing of the audio component it was observed that some of the collected user audio data, involving 23 subjects in total, contained extraneous environmental noise which could interfere with the recognition process.

In this case, it was necessary to preprocess the audio data of these 23 participants in order to improve the signal quality. The Figure 5-3 shows an example, taken from one of the users of the system. In the upper part, a noisy signal is given, and in the lower part, a cleaned one. A simple wavelet of the sym4 type is used; the result of cleaning is clearly visible. During noise cleaning, small threshold values were used in the range from 0.05 up to 0.3 to avoid information loss.

Threshold	**Number of cleaned audio files**
0.05	0
0.1	180
0.15	290
0.2	94
0.25	91

Table 5.6: Effect of changing threshold value for noise frequencies on cleaning data of participants

From the Table 5.6 it is seen that the number of files cleaned is relatively small in comparison with whole available data volume. For the remaining files, raising the noise level threshold was considered inappropriate due to the potential loss of important parts of the signal carrying information about the user. It was concluded that additional preprocessing would be required for the remaining cases.

After receiving the results of denoising audio files, the signal components were segmented into empty and useful frames. During observation of VAD algorithm performance, it was determined that the selected basic parameters for signal separation satisfactorily dealt with the SpeakingFaces dataset audio files. As an example, the Figure 5-4 shows the clearing of the entire signal from the intervals containing silence, where the signal is located on top after removing noise, and on the bottom is the result of applying the VAD algorithm. As seen from the diagram, areas with the lowest values of signal frequencies are detected and removed from recording of the final data to file after processing.

In general, the preprocessing of audio files based on the proposed architecture components of Zhang et al. [59] before building a model for the SpeakingFaces target

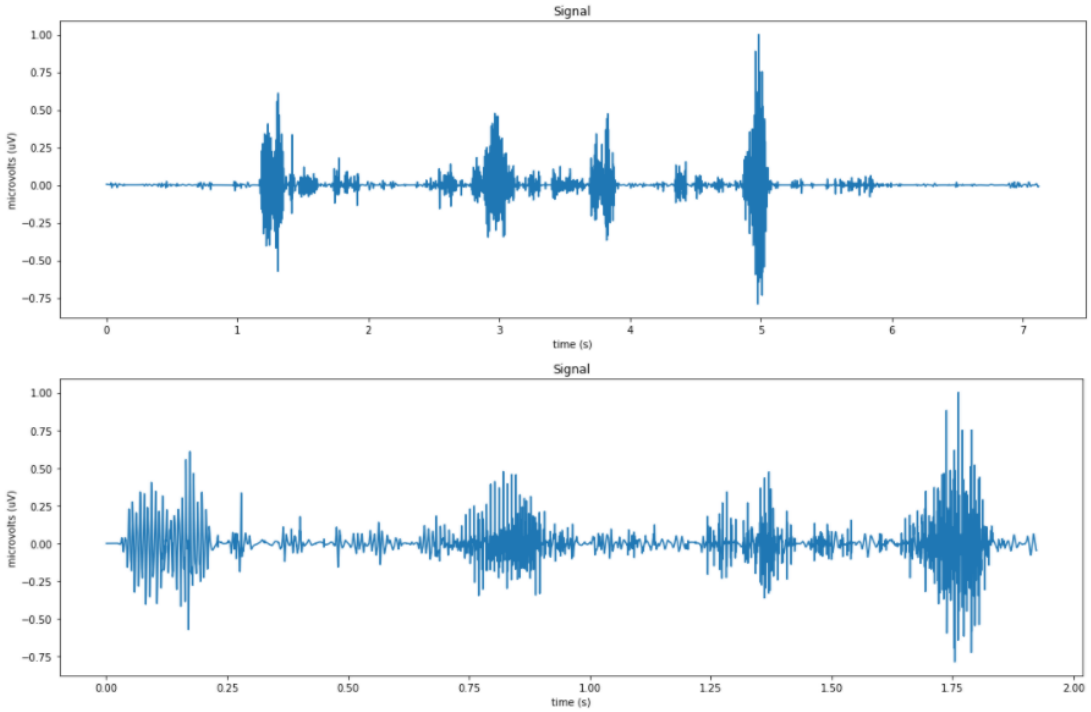


Figure 5-4: Example of VAD application on audio file

Architecture	**Accuracy**
With preprocessing	65.19%
Without preprocessing	77.55%

Table 5.7: Results of user recognition depending on preprocessing presence

dataset did not improve the quality of user recognition, since some of the data that could carry potentially important voice features was removed. From the Table 5.7 of test results of the best model, you can see that with the application of techniques to improve the signal quality, the accuracy of identifying the speaker of the system drops by 12.36%. From this we can conclude that the preprocessing techniques used in the work [59] are not relevant for processing this dataset and that higher recognition accuracy could be achieved if more attention was paid to taking into account external factors during the data collection phase of the dataset such as wind noise, sounds of equipment, the distance of speaker to the microphone and etc.

5.2 Multimodal Biometric System

In the scope of this thesis work, two fusion systems were created and validated. Due to the complexity of the multimodal system and hardware parameters, only frontal faces of users was validated on the multimodal authentication system.

Table 5.8 displays the performance of the threshold based fusion system proposed by Zhang et al. [59] on the frontal faces SpeakingFaces subset with various values for thresholds t_sum and t_prod. Their fusion model takes results from two unimodal components as well as the username of the user. As it can be observed from the table 5.8, increasing t_prod threshold with particular value of the t_sum may not have any effect on the TAR or may diminish it, while increasing t_sum threshold resulted in higher accuracy.

Parameter	**System setup and results**					
a	0.5	0.5	0.5	**0.5**	**0.5**	**0.5**
t_sum	0.01	0.01	0.01	**0.05**	**0.1**	**0.1**
t_prod	0.01	0.05	0.1	**0.01**	**0.01**	**0.1**
TAR	61.91%	65.31%	65.31%	**92.52%**	**98.3%**	**97.96%**
FRR	38.09%	34.69%	34.69%	**7.48%**	**1.7%**	**2.04%**

Table 5.8: Performance of the threshold based fusion score

Zhang et al. [59] achieved 100% TAR on the XJTU database and our threshold based model achieved the highest achieved result yet of 98.3%. Since thresholds are used for authentication purposes, it is convenient to set them as minimum as possible. Three highlighted model parameters with (0.05; 0.01), (0.1; 0.01) and (0.1; 0.1) value pairs for (t_sum; t_prod) parameters with more accurate authentication rate were chosen as best models for threshold based fusion model. The splitting criteria parameter 'a' was set as 0.5 in order to treat the results of both unimodal systems equally.

Table 5.9 illustrates the performance of the ranking based fusion system on the frontal faces SpeakingFaces subset. Ranking based fusion system did not require to enter the username; on the contrary, it checks whether the topN users from voice and face are identical and the closest user is detected. In table 5.9 FRR indicates the

percentage of users who should be granted access but were rejected by the system and the False Detection Rate (FDR) was counted as the percentage of occurrences when one user was incorrectly identified as different one.

Rank	**TOP 5**	**TOP7**	**TOP10**
TAR	67.35%	72.79%	76.53%
FRR	29.59%	19.39%	8.5%
FDR	3.06%	7.82%	14.97%

Table 5.9: Performance of the implemented ranking fusion score

As it can be observed from results of the TOP 5, TOP 7 and TOP 10 ranking fusion scores, smaller number of the similar users checked from the database led to the lower accuracy of the authentication rate, but also in the modest rate of False Detection Rate.

All results from multimodal biometric systems were obtained by testing 3 frontal faces of 98 subjects in database with random 3 audio files of the user. In order to verify the effect of all parameters, seed was set and for each experiment same audio files within same faces was validated. However, due to randomness of audio files TAR value may fluctuates within 2-3% from the mentioned values.

5.3 Validation on External Datasets

After creating and assessing the performance of the unimodal components on the SpeakingFaces subset, the systems were validated using the Georgia Tech Face Database and the TIMIT dataset.

5.3.1 Georgia Tech Face Database *(AI)*

The Georgia Tech Face Database was assessed on distinct parameters compared to the SpeakingFaces dataset; it used the same architecture, but with customized parameters. The major reason for this is, as noted, the background noise that interferes with the facial detection when using the best parameters for the SpeakingFaces subset. The best obtained parameters for the Georgia Face Tech database were 1.001 scaling

factor with 8 neighbors, the minimum size of the face image 140*140 and City-block metric. Moreover, some preprocessing procedures were required to diminish the effect of the noisy background (see Appendix A).

The default template of Haar cascades with best defined parameters resulted in 97.7% accurate recognition face rate, while Zhang et al. [60] achieved 97.2% of recognition rate. Table 5.10 displays the best results achieved for the both SpeakingFaces and Georgia Tech Face datasets. The state-of-art approaches from Table 2.1 achieved 100% as a best result and 70.71% top 5 result for the Georgia Tech Face database.

Dataset	**Presence within TOP 5**	**TAR**	**Average recognition rate**
Implemented architecture results			
SpeakingFaces	81.29%	71.09%	99%
Georgia Tech	80%	69.33%	97.7%
Zhang et al. [59] results			
XJTU	-	91.64%	96.6%
Georgia Tech [60]	-	90.14%	97.2%

Table 5.10: Comparison of approaches' results with validation Georgia Tech Face Database

Our model achieved 69.33% TAR for the determined parameters and 80% presence within top 5 candidates. The performance of the current unimodal component is superficial when compared with top 5 accuracy of the state-of-the-art approach. The model was also tested with Euclidian metric and similarly to the SpeakingFaces dataset, the TAR declined to 52% and top 5 accuracy declined to 67.33%. The replicated model has less accuracy on the validation Georgia Tech Face Database due to the different approach in the user verification part. The prior unimodal component of Zhang et al. [59] and [60] provides to the system the username and by selecting the threshold decides whether a user is authorized or not.

5.3.2 TIMIT *(AK)*

To check the reliability of the results of the system based on the best parameters of the model, the TIMIT dataset was chosen, which was also used during the validation of the architecture from the replicated work [59]. The quality of the data in the validation

dataset differs from the SpeakingFaces dataset in the absence of interference in data collection. When listening to audio recordings, the speech of the participant is clearly audible.

The main results were obtained on a model with the following hyperparameters: 150 components, 50 EM iterations, 'tied' covariance matrix and 1 initialization, and reflected in Table 5.11.

Dataset	**Presence within TOP 5**	**TAR**	**FRR**
Implemented architecture results			
SpeakingFaces	92.18%	77.55%	22.45%
TIMIT	99.78%	96.5%	3.5%
Zhang et al. [59] results			
XJTU	-	89.02%	10.98%
TIMIT [61]	-	92.65%	7.35%

Table 5.11: Comparison of approaches' results with validation dataset TIMIT

As a result of testing the model on the TIMIT dataset, the recognition accuracy reached 96.5%, which is close to the value obtained using the same architecture in the work of Zhang et al. [59], where the TAR value is equal to 92.65%. The difference of 3.85% may occur due to the absence of a description of the parameter values in the prior work; without these values it was difficult to fully reproduce their results. In general, given the high corresponding values of the user recognition accuracy on the validated dataset, it can be concluded that the replication of the voice module was successful.

Chapter 6

Spoofing Attempts and System Response

This chapter analyzes the effect of four attacking scenarios on our multimodal authentication system under two different fusion score schemes. The four scenarios of no spoofing, face-only spoofing, voice-only spoofing and both biometrics spoofing were used to compare the security level of the fusion scores. For the last case of combined spoofing, several imitation procedures were conducted.

Since our multimodal biometric system consists of unimodal components and combines them using particular thresholds and other techniques, it is expected that higher value thresholds and scores will lead to a lower security level of the system and higher FAR values.

6.1 Biometrics Mimicry

While checking the system for vulnerabilities, several simulations were carried out of instances when a person from outside of an organization wanted to gain unauthorized access through the spoofing of biometrics of another user. To replicate the biometrics of users, the data of the system intruders were used. A search was made for the people most similar to the intruders among all users of the system, as a result of which 11 pairs of participants with visually common features were identified.

To test the authentication system for attempts to spoof the biometrics of the user's face, it was required to obtain images with combined facial features of both the intruder and the existing user. The free online service 3Dthis.com was used to create morphed images; the service provides a variety of applications related to the creation of 3D illustrations and animations, including the Face Morph application. Both images belonging to the attacker and the user are loaded to the application through an interactive interface. Next, the images are aligned according to key features of the face, including aligning image sizes and overlaying features exactly on top of each other. Based on the actions taken, the application creates a morphed image, which is available for free download. These steps were performed for each pair of images of selected similar users and intruders. Figures 6-1 - 6-3 show the original images before morphing and the resulting image after processing by the application.

Figure 6-1: Original image of intruder

Figure 6-2: Original image of system user

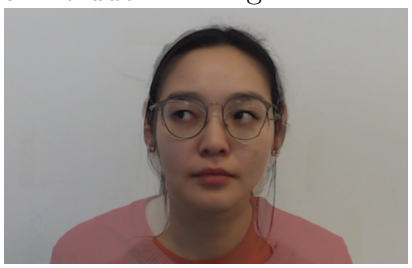


Figure 6-3: Result of morphing original images

A slightly different approach was used to simulate voice biometrics. In this case, based on the audio recording of the system user, a new artificial speech signal was created using the work of Yuxuan Wang et al. [55] who created the Tacotron system that synthesizes human speech using a text phrase directly from the user's audio signal. The implementation of the practical part of this work is in the public access

on the GitHub platform, which made it possible to run the code on audio files of the same selected users of the system as during image morphing for further testing of spoofing of both biometrics together. The code takes the user's audio recording, where further processing is done. It is also possible to record a voice of any length on site using a microphone. Next, a phrase is selected that will be present in the synthesized audio recording, and the signal itself is generated.

6.2 No Spoofing

The case of no spoofing represents the situation when an unauthorized person tries to get an access to the system with their own biometrics. Since the threshold based fusion score requires the username of the user, two experiments were conducted to verify the data confidentiality under this multimodal system. First the experiment is conducted when each intruder tries to authenticate with the username of an authorized user in the database and a second experiment was conducted for 11 intruders from Section 6.1, which attempts to authenticate in the system as a user who looks similar to the attacker.

Parameter	**Model performance**		
a	0.5	0.5	0.5
t_sum	0.05	0.1	0.1
t_prod	0.01	0.01	0.1
TRR	86.97%	64.67%	64.67%
FAR	11.03%	33.33%	33.33%

Table 6.1: Performance of the threshold based fusion score on 7,350 attacking attempts

Parameter	**Model performance**		
a	0.5	0.5	0.5
t_sum	0.05	0.1	0.1
t_prod	0.01	0.01	0.1
TRR	76.56%	57.81%	57.81%
FAR	23.44%	42.19%	42.19%

Table 6.2: Performance of the threshold based fusion score on 64 attacking attempts

Table 6.1 and table 6.2 display results of the system security assessments on the first and second experiment respectively. Table 6.1 represents the case when all 13 intruders try to authenticate as each of the 98 authorised users by providing to the system the username of an authorized user, while table 6.2 analyses the scenario when 11 intruders provide the username of a similar-looking authorized user. As it can be observed from the tables, the targeted attacks when an intruder attempted to enter into the system by providing similar looking authorized users login are more successful than attempts when the attacker knows all authorized usernames in the database and tries to enter the system with each of them.

Parameter	**Model performance**		
Rank	**TOP5**	**TOP7**	**TOP10**
TRR	72%	58.67%	40%
FAR	28%	41.33%	60%

Table 6.3: Performance of the ranking based fusion score on 75 attacking attempts

Table 6.3 shows the performance of the based fusion score on 75 attacks performed by 13 intruders. For ranking based fusion score the attacker provides only its biometrics and the system decides whether it should grant access and under which authorized users's credentials.

As it was expected, the higher thresholds and ranking number lead to the greater vulnerability of the system under external attacks. Overall the performance of the threshold based system is better for the no spoofing scenario. However, for the complete security analysis of the system, the dynamics of FAR value changing under other spoofing scenarios for both fusion systems should be reviewed.

6.3 Face-only Spoofing

The face-only spoofing scenario is the case when 11 intruders are using own voice biometric and face image of similar looking authorized user. Both fusion scores results were obtained on 11 attacks by assuming that intruder has only one image of the real user from the system.

Parameter	Model performance		
a	0.5	0.5	0.5
t_sum	0.05	0.1	0.1
t_prod	0.01	0.01	0.1
TRR	81.82%	72.73%	72.73%
FAR	18.18%	27.27%	27.27%

Table 6.4: Performance of the threshold based fusion score on face spoofing scenario

Parameter	Model performance		
Rank	**TOP5**	**TOP7**	**TOP10**
TRR	72.73%	54.55%	36.36%
FAR	27.27%	45.45%	63.64%

Table 6.5: Performance of the ranking based fusion score on face spoofing scenario

Table 6.4 and table 6.5 illustrate threshold based fusion and ranking based fusion results under face only spoofing attack scenario respectively. The two multimodal systems were able to maintain almost the same level of TRR comparing to results from the no spoofing scenario. As a result, both fusion scores shows the resilience to face spoofing attacks.

6.4 Voice-only Spoofing

Similarly to the face only spoofing scenario, the voice spoofing scenario uses the intruder's image with audio of the similar looking user. The audio recordings were created from the second session of each user's data and did not contain the same phrases or commands which were stored in the database. For validation of the biometric authentication system confidentiality 64 tests were performed: 11 intruders had several voice recordings of the authorized user.

Parameter	Model performance		
a	0.5	0.5	0.5
t_sum	0.05	0.1	0.1
t_prod	0.01	0.01	0.1
TRR	15.62%	4.69%	4.69%
FAR	84.38%	95.31%	95.31%

Table 6.6: Performance of the threshold based fusion score on voice spoofing scenario

Table 6.6 demonstrates the threshold based fusion system response under the voice only spoofing scenario. It can be seen that model is dramatically vulnerable under voice attacks case and the FAR value for each setup is higher than 83%. As a result, in more than 4 out 5 cases when an intruder has voice recordings of the authorized user the attack on the system is successful.

On the other hand, the performance of the ranking based fusion system which can be reviewed from the table 6.7 is improved a little bit compared to the no spoofing and face only spoofing scenarios. This multimodal system demonstrates sufficient resilience toward voice based attacks.

Parameter	**Model performance**		
Rank	**TOP5**	**TOP7**	**TOP10**
TRR	75%	62.5%	40.63%
FAR	25%	37.5%	59.37%

Table 6.7: Performance of the ranking based fusion score on voice spoofing scenario

By comparing the level of data security under two described biometric multimodal system, it can be concluded that the ranking based fusion model has a minimal vulnerability under single biometric system attacks.

6.5 Both Biometrics Spoofing

For both biometrics spoofing scenario authentication systems are attacked with morphed face and synthesized voice data.

Parameter	**Model performance**		
a	0.5	0.5	0.5
t_sum	0.05	0.1	0.1
t_prod	0.01	0.01	0.1
TRR	100%	72.73%	72.73%
FAR	0%	27.27%	27.27%

Table 6.8: Performance of the threshold based fusion score under morphed data attacks

Table 6.8 and table 6.9 represents the authentication model response under morphed intruder's data. By comparing obtained results with no spoofing scenario out-

Parameter	Model performance		
Rank	**TOP5**	**TOP7**	**TOP10**
TRR	100%	72.73%	54.55%
FAR	0%	27.27%	45.45%

Table 6.9: Performance of the ranking based fusion score under morphed data attacks

come, it can be clearly observed that both systems provide better security on the morphed results rather than on original intruder biometric data.

In order to determine the reason of the great model resistance under morphing attack, additional experiments were conducted. For assessing quality of the imitated biometric data following setup was used: face morph with one stolen from authorized user voice data and voice imitation with one stolen from authorized user image.

(t_sum, t_prod) pairs	(0.05; 0.01)		(0.1; 0.01)		(0.1; 0.1)	
Results	Face	Voice	Face	Voice	Face	Voice
TRR	36.36%	100%	18.18%	72.73%	18.18%	72.73%
FAR	63.64%	0%	81.82%	27.27%	81.82%	27.27%

Table 6.10: Performance of the threshold based fusion score under single generated data attack

Results	TOP 5		TOP 7		TOP 10	
	Face	Voice	Face	Voice	Face	Voice
TRR	72.73%	90.91%	18.18%	63.64%	0%	27.27%
FAR	27.27%	9.09%	81.82%	36.36%	100%	72.73%

Table 6.11: Performance of the ranking based fusion score under single generated data attack

From Table 6.10 and Table 6.11, it can be concluded that the quality of face morph is sufficient to identify that morphed image is in TOP 10 closest images to the real system user. The vulnerability of the voice synthesized audio data is a result of the lower authentication rate for the real system users from which data was chosen. Ranking based fusion score results concluded that morph data in almost 7 out of 10 cases can be identified as TOP 10 closest to the some authorized user. On the other hand, the same voice synthesized audio data was rejected 7 out of 10 times with the threshold based fusion system which implies that a picked to imitation user can be misidentified as another user in the system.

As a conclusion, the dynamics of system responses under several spoofing attack scenarios varies for ranking based fusion system and threshold based fusion system. While both systems maintain almost stable resilience under non-spoofing and face spoofing cases, the threshold based fusion system demonstrated high vulnerability on the voice based spoofing attacks and contrarily the ranking based fusion score within top 10 closest to the system users provided a significantly high level of FAR.

Chapter 7

System Analysis

This chapter provides additional analysis on the outcomes from the prior chapters. The scrutiny provides details of the confidentiality, integrity and availability of the designed multimodal biometric authentication system. Each category evaluates the system responses within various experiments covered in previous sections. In addition, discussion on some limitations of the model and computing resources is presented as a part of the system analysis.

7.1 Confidentiality

The major concern for storing personal data on the any type of the data storage is the security and privacy of the user's data. The confidentiality of the authentication system is the crucial factor for implementing it in the real world. Contemporary methods of providing sufficient level of security and privacy of the data include plentiful encryption algorithms from the cryptography field.

The confidentiality investigation of our created multimodal biometric authentication model relies on the data provided after spoofing attempts by external attackers. As it was mentioned in Ch.6, both the threshold based fusion and ranking based fusion models response were analyzed under four different attack scenarios.

Fig. 7-1 represents the dynamics of the system response under different type of attacks. From outcomes provided in Ch.6 two systems were identified as the

most resilient under intruder attempts: top 5 system from ranking based fusion systems and threshold based model with following parameters: a=0.5, t_sum=0.05 and t_prod=0.01.

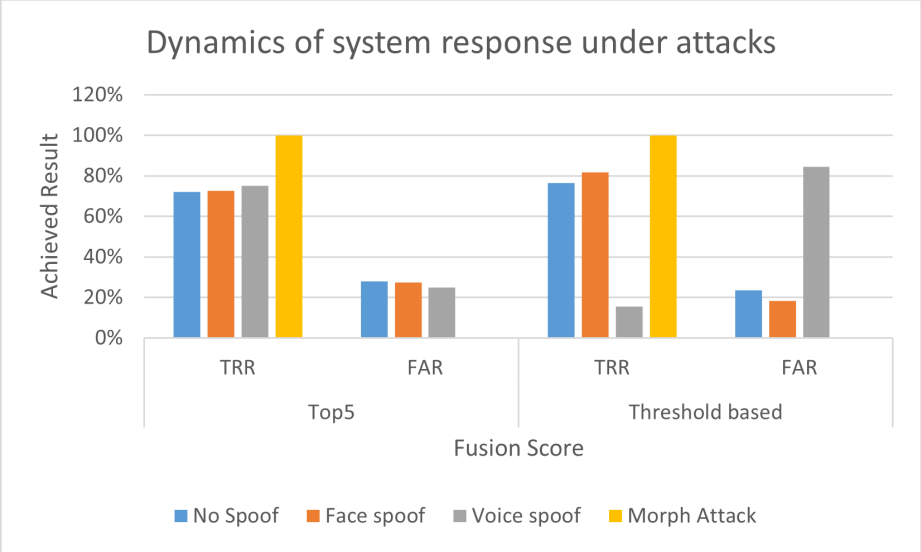


Figure 7-1: Dynamics of changing TRR and FAR value for two fusion score schemes under different intruders attacks

It was expected that systems with lower thresholds and lower ranking number will provide higher security level of the biometric authentication system. From Fig. 7-1 we can derive that top5 system provides steady level of the security for the system with TAR value higher 72% under all presented attacks. On the other hand, threshold based maintains greater level of data protection but fails to deliver same level of security under voice spoofing scenario.

By comparing results from both systems, it is reasonable to prefer the system with stable level of the security under different circumstances to the system with a little bit higher protection level for some potential attacks, but inadequate vulnerability under other attacks. In addition, FDR value was introduced in the Ch.5 for ranking based fusion model, which analyzed the probability that system will misclassify user as another one and provide an access to the sensitive data of another user. Top 5 ranking system maintain FDR rate lower than 4% for the 98 authorized users. For the further researches it is recommended to compare system responses under other fusion systems and increase the variety of the attacks on the system as well

as FDR analysis on the threshold based system, which may be in case when user mistypes its username, to provide the full report on system behaviour. Moreover, the security level of the system can be improved by adding additional model components such as liveliness detection, and can be validated on the video type data from the SpeakingFaces Dataset.

7.2 Integrity

Another important factor for the data storage is system integrity. This factor provides an adequate protection to the unauthorized data modification procedures (deletion, insertion and corruption) from human or system itself. The possible intentional and unintentional human intervention into the system and access to the restricted sensitive data was discussed in Section 7.1. Both fusion systems provide adequate level of security for some scenarios; however, the possibility of unauthorized data modification is still present in both models.

The scope of thesis work does not cover possible data changes from the system or storage itself. All models and experiments were conducted on the Google Cloud computational services and there were no any possibility to check whether some specific system procedures (maintain nights, server rebooting, etc.) affect the vulnerability of stored data. For the studies with their own data storage component, server or hardware - it is recommended to analyze the possible data modifications due to storage instability.

7.3 Availability

The final significant parameter in the system analysis is availability. Potential users are evaluating authentication models mostly on the reliability factor, accuracy of the authentication process and timely response of the system.

7.3.1 Reliability

Reliability factor is the primary parameter in the process of creation the authentication model. Inadequate system operation or constant hardware crashes will lead to additional expenses for modification of the model, repair works or full replacement of hardware and model.

The reliability scrutiny of the implemented system on the SpeakingFaces subset was conducted during the implementation stage. Some parameter values of the model led to crashes of the authentication process. The unstable system performance was caused by the unimodal face recognition component. Two major variables: number of neighbors and scaling factor - provoked multiple crashes of the system. For the best model scaling factor higher than 1.01 and number of neighbors higher than 18 caused failure to deliver stable work of the authentication model. Various range of values for both parameters: 1.015, 1.02 and 1.05 for scaling factor and 20, 26 and 34 number of neighbors - were infeasible for model setups due to system collapses.

The hardware reliability analysis was implemented during model optimization stage. Google Cloud computational services offered three processor types: CPU, GPU and TPU. After several experiments it was detected that GPU processor can not provide sufficient level of support for system operations. Probably due to parallel computing properties of GPU cores, the RAM storage capacity filled out quickly and resulted in the system crash.

Future studies may add some preprocessing procedures in order to determine the potential value range for the parameters of the biometric authentication system for SpeakingFaces dataset. Moreover, new hardware systems may provide higher RAM capacity rather than free available cloud services and the time-cost analysis of the GPU type processor can be conducted.

7.3.2 Time-Cost and Accuracy Analysis

Time-Cost Analysis

Modern technologies and researchers are working toward faster system response on user's request. Time-cost analysis with accuracy results provides to the user a better understanding of model performance. For negligible difference in the accuracy of the system, a user will prefer model with quick response and vice versa. For the implemented biometric authentication model time-cost analysis was conducted on unimodal component level as well as on the multimodal systems.

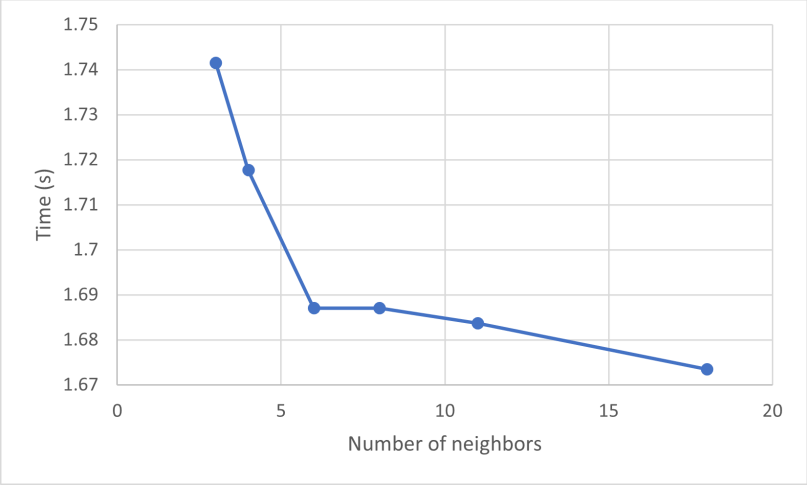


Figure 7-2: Effect of the number of neighbors on the single user authentication

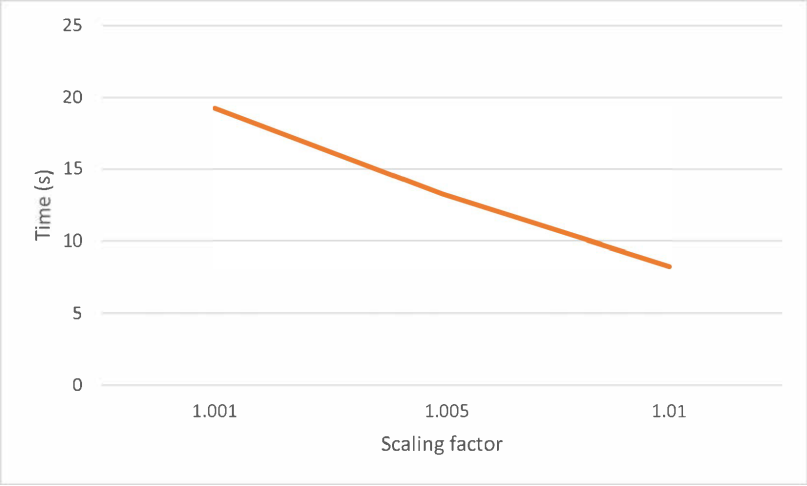


Figure 7-3: Effect of the scaling factor on the single user authentication

Figure 7-2 and Figure 7-3 graphically illustrates the correlation between time response and parameters of the unimodal face recognition component. The growth in the number of neighbors from 3 to 18 resulted in 70ms faster face recognition and increase of the scaling factor from 1.001 to 1.01 led to 2 times faster response from the face component. The best model with 1.01 scaling factor and 18 number of neighbors managed almost 11s reduction of the authentication process.

Figure 7-4 represents a time cost analysis of using different processor types for both training and testing of the unimodal face component. As it can be seen the CPU processor type works 4 times faster in both processes comparing to TPU core. As for the unimodal voice component, there were restrictions due to limited computing and data resources when extracting the main signal features and assembling the classification model. For that reason, all the code was run in Google Colaboratory environment, which provided default amount of computing resources sufficient not to use the GPU and TPU hardware accelerators.

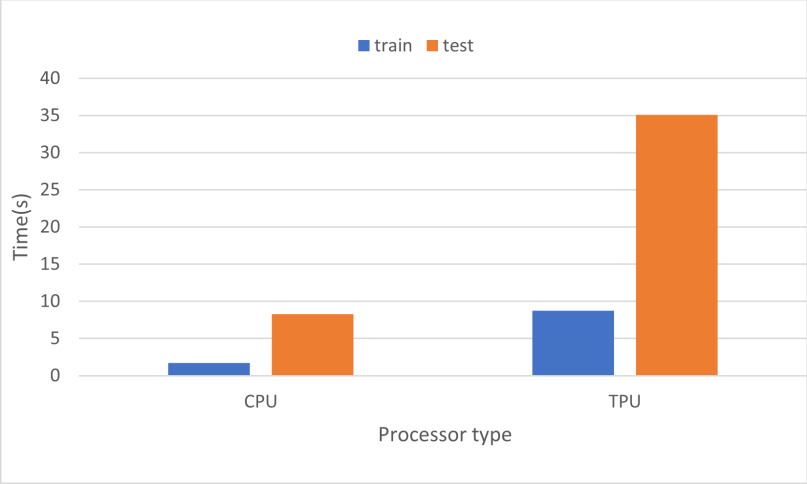


Figure 7-4: Effect of the processors cores type on the single user authentication

Regarding the influence of the processes of training and testing the user's speech recognition model on the overall performance of the system, it was mentioned in the results section 5.1.2 that with an increase in the number of model parameters, the time spent on training and testing grows in direct proportion. Even if the usage of larger number of distributions per model increases recognition efficiency, it also

increases the dimension of the decision-making space and processing time.

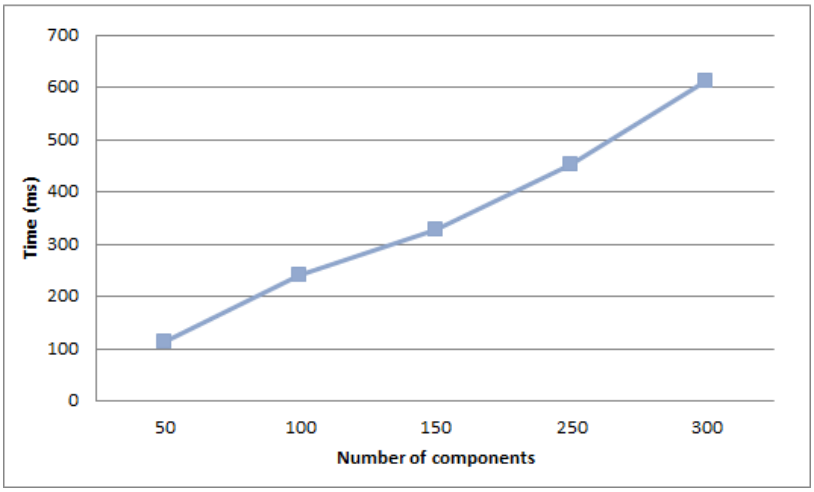


Figure 7-5: User authentication time dependence on number of components model parameter

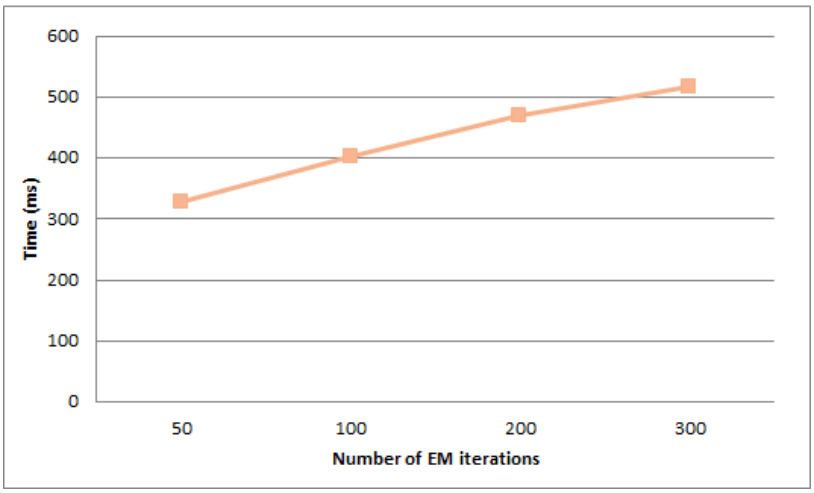


Figure 7-6: User authentication time dependence on number of EM iterations model parameter

The Figure 7-5 shows a graph of the change in the duration of user authentication after reading an audio file depending on the parameter of the number of model components. Here it can be seen that the time spent on checking the user grows almost linearly. Given that there was no significant difference in the accuracy of the user recognition after 150 components, for this case it was relevant not to increase the number of components above this value in order to avoid wasting time. In addi-

tion, the second parameter, the number of EM iterations of the model, added time to authenticate the user. The Figure 7-6 shows the dependence of the time for testing the system on one user with the same number of components (150). As can be seen from the graph, with an increase in the number of iterations, the time spent on user authentication also grows. Therefore, as the final value of the EM iterations parameter, the value equal to 50 was chosen. It takes about 2.5 hours to train the model with the best obtained accuracy and optimal parameters, with testing on the same parameters in 1 hour 4 minutes. Thus, it takes 328 milliseconds to authenticate one user.

Figure 7-7 displays the time difference of single user authentication process for two fusion systems: ranking based and threshold based. For time-cost scrutiny of multimodal system were used best parameters form unimodal components which provide quicker response and CPU processor type, which boosts model performance by delivering faster response. Three ranking based fusion setups: top 5, top 7 and top 10 - in average generates response in 8.9s and variance of this indicator is almost 0.1s; while for threshold based systems with (t_sum; t_prod) pairs: (0.05; 0.01), (0.1; 0.01), (0.1; 0.1) - time response ranges between 8.8s and 10.4s with average time response - 9.6s.

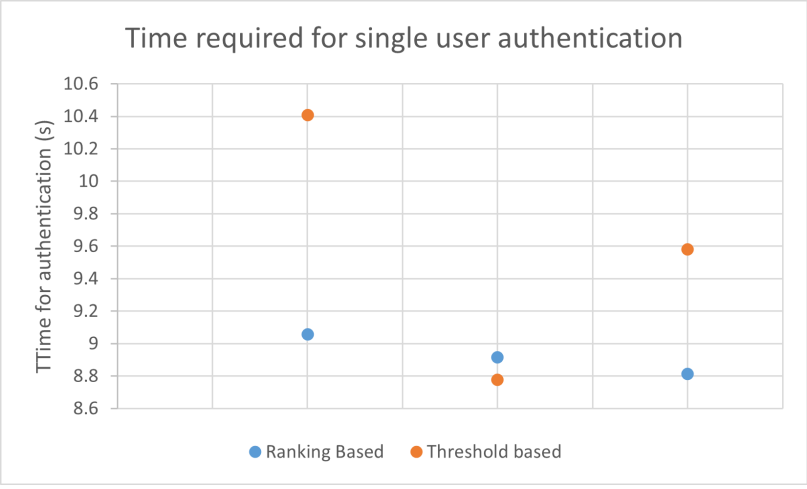


Figure 7-7: Single user authentication time analysis between two fusion systems

Accuracy

Accuracy is the crucial indicator of proper authentication system and the majority number of studies are focused mostly on the improvement of the accuracy score. Since, the accuracy results were provided in the Ch.5, this section presents some discussion on the comparison between two fusion models and limitations of resources which may resulted in the lower accuracy of the implemented system.

Figure 7-8 demonstrates the difference of TAR in the achieved best 3 models from two fusion systems: ranking based and threshold based. In general, the threshold based system provides better accuracy in the authentication process rather than the ranking based system. The ranking based fusion system provides lower results due to moderate accuracies of the unimodal components, and can be improved by achieving better results from both unimodal systems. Nevertheless, by summing up system analysis for both systems we can conclude that more precise in accuracy threshold based fusion system is on average on 0.7s slower than the fusion based model and can led to high vulnerability under some type of intruder attacks.

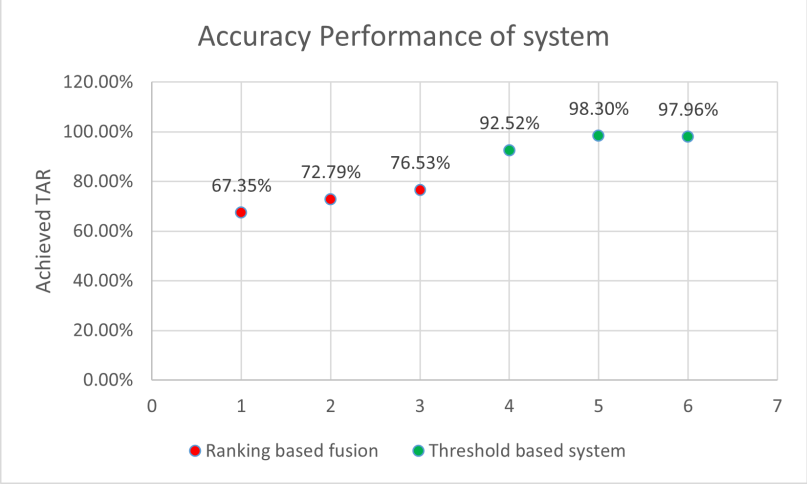


Figure 7-8: Achieved TAR for different setups of multimodal systems

In addition, to build an adequate biometric authentication system for the user reference base, it is necessary that it contains image samples with different person visuals and speech signals with duration of tens and even hundreds of hours, which does not correspond to the total volume of collected data from the tested dataset.

Thus, building a full-fledged authentication system with sufficient user recognition accuracy will require a significantly larger amount of available resources.

Chapter 8

Conclusion

The aim of this thesis was to recreate and extend the architecture of a multimodal biometric authentication system based on recently published work by Zhang et al. [59] and examine the confidentiality, integrity, availability and performance of the system developed using a subset of the novel new large-scale SpeakingFaces dataset and validated against two open-source datasets.

The whole work was presented in eight chapters.

Introduction chapter provided historical overview on the usage of the biometric data and the motivation behind implementation of the multimodal biometric systems. Moreover, thesis objectives and goals were discussed as well the detailed structure of thesis document.

Chapter two reflected the literature review on the recent studies conducted in the biometric authentication field. There were discussed methods that used both physical and behavioural biometric data for implementing unimodal and multimodal authentication system. Some the state-of-art results as well as their approaches was provided for well-known datasets for unimodal face and voice components. The detailed description of the published work by Zhang et al. [59] serves as a foundation to the implemented model as part of the scope of thesis work. In addition, the review of existing studies which focused on the confidentiality and optimization aspects of the system were examined.

Chapter three was dedicated to introduction of three datasets: SpeakingFaces,

TIMIT and Georgia Tech Face. The data collection and initial preprocessing on the data done by publishers was described as well as some limitations of dataset discovered during data analysis process. Supplementary data related procedures from data preparation stage and selection process for the SpeakingFaces subset were fully described.

The methodology of implemented biometric authentication system was covered in chapter four. Unimodal face and voice component and fusion score calculation architectures were presented with description of each sub-module component and method. Moreover, the limitation of prior work [59] and some adjustments to each module was discussed and all parameter sets for experiment phase were determined.

Chapter five displayed the results obtained from different simulation of the created model. Initially, the performance of the unimodal components were analyzed and the best parameters in terms of accuracy of the model were established. Two fusion score calculations: frequency based and threshold based fusion - were conducted on the best parameters obtained from the unimodal components. Additional analysis of the thresholds and rank on the model were investigated and the highest TAR score of 98.3% was achieved with threshold based fusion model on the subset of SpeakingFaces dataset. Furthermore, the model performance was validated on the TIMIT and Georgia Tech Face datasets. Unimodal voice component's results on validation TIMIT dataset are in line with state-of-the-art results, while unimodal face system accuracy and presence within TOP5 results were close to some state-of-the-art achievements.

In chapter six the system response under different scenario of intruder's attack was assessed. Before analyzing the FAR and TRR scores of the system, the process of creation morphed face image and voice synthesis were explained. Three best models of two fusion scores were evaluated under four spoofing techniques scenarios: no spoofing, face-only spoofing, voice-only spoofing and both biometric data spoofing. The ranking based fusion score models provided stable resilience under all scenarios, while threshold based fusion systems demonstrated high vulnerability under voice-only spoofing attack. However, the impact of the single synthetically obtained data :

morphed face and voice synthesis - was evaluated on the all models. The quality of the generated data was proved by investigating the response of ranking based systems. Both generated data type was identified as TOP7 and TOP10 closest biometric data to the existed in the database user.

Chapter seven provided the full scrutiny on the confidentiality, integrity and availability of the system within scope of thesis work. The system confidentiality and integrity analyses were based on system response under spoofing attacks from chapter 6. The dynamics of system response under spoofing attacks was compared between the best models from two fusion score schemes. TOP5 ranking model maintained TRR value to be higher than 72% for all cases and threshold pair (t_sum=0.05; t_prod=0.01) of threshold based model results were higher than 75% except for voice-only spoofing attack, which demonstrated near to 85% FAR value. The system availability analysis were covered within system reliability factor and time-cost and accuracy scrutinises. Findings during the model implementation phase were described in reliability phase, while time-cost analysis provided the dependency of the time of user authentication process on the different system component. Moreover, the additional analysis on the hardware component on the time of system operation was discussed. As a result, the threshold based fusion models were able to provide higher TAR score between 92.52% and 98.3% but in average they working slower than ranking based fusion models.

This paper goals and aims were achieved during all stages described in the prior chapter. The implemented multimodal biometric authentication system provided sufficient value for TAR with highest 98.3%, which compensates the moderate performance of the unimodal components. However, the confidence in the biometric models was undermined after attacking the system and obtaining the situation when 4 out of 5 attacks from unauthorized person were successful. In addition, some limitation of the current work and recommendations for the further researches are provided.

8.1 Limitations

There are several limitations to both the prior work and replicated system that were discovered during the project, and described in the prior chapters.

8.1.1 Hardware and Computational Resources

As it was discussed, the computational demands of our model significantly exceed the capacity of higher-end desktop systems, even those configured with dedicated GPU cards. After consideration of options, we determined that the model would be better supported under new custom-designed architectures for deep-learning, such as the NVIDIA DGX systems.

In this regard, the best option appeared to be the Google Colaboratory Cloud service. which provides access to DGX computational resources and cloud storage. The service worked well, in general, but nevertheless was sub-optimal due to system quota constraints; model training and testing took considerable amounts of time to run, thus imposing limits on elements such as fine-tuning system parameters by running multiple sessions across a wider range of configurations.

Under Google Colab, resource allocation is performed by the service itself, and may vary over sessions, thus performance and stability of the system were uneven. Further, the scaling of the system within the Google Colab environment remains as an open question, in terms of whether a more complex model would be supported.

On this point, it is our assessment that while Google Colab was a good environment for initial development and testing of the prototype of the model, it is not clear that it would serve as well for elaboration of the model nor deployment of the system for real-time applications.

8.1.2 Dataset

The SpeakingFaces dataset used as the main dataset for the model implementation and performance validation also contains thermal images of the participants. We considered integrating the thermal images into our model, however, we observed from

the literature that existing data sets typically contained a wider range of user states, either environmental or emotional, whereas SpeakingFaces was limited to two sessions, with less consideration of the the environmental or emotional range of the subjects. For a more robust user authentication system, the thermal data should be collected from each state of the user which may vary due to weather conditions and mental state of the person.

In addition, there were some missing files and some data irregularities identified for couple of the users. Slight blurred or cropped face and background noises in the audio data affected the model performance for some small part of users.

8.1.3 Model

The study published by Zhang et al. [59] which serves as the foundation for the implemented model was missing some crucial parameters regarding the initial setup of the model and relevant threshold values. As a result, some procedures such as image preprocessing for the face module and model parameters slightly differ from the work by Zhang et al [59]. For unimodal component architecture, the prior works by Zhang et al. [60], [61] were used to specify some parts of the models. As a part of face recognition, there are limitation in the number of Haar cascade templates available for the profile faces of users, which affected the face detection rate and TAR values.

Due to absence of the thresholds for fusion score, a range of thresholds were tested, and ranking based fusion scores were implemented. However, the confidentiality analysis may not reflect the accurate results on the proposed model by Zhang et al. [59] and system response can differ due to the omission of threshold parameters.

8.1.4 Time

The timeline for thesis completion imposed a limit on the range of scenarios that could be tested, and on the total number of test subjects used for training and testing. As an example, the prototype for facial recognition was implemented only for frontal face

profiles.

8.2 Future Works

For the further researches it is recommended to conduct several changes in the our proposed approach. One suggestion is to implement other unimodal face and voice modules to analyze which approach provides better performance of the model. Moreover, it is suggested to change the range of the listed parameters of the model components and examine a wider range of tuning parameters for accuracy of the system.

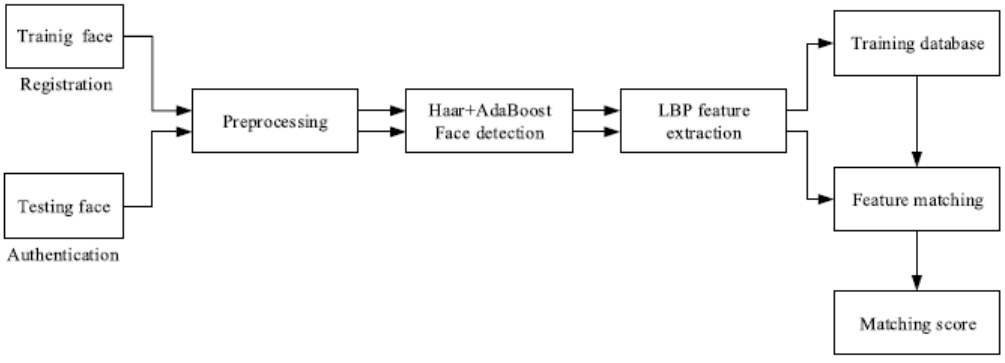
In order to more comprehensively assess the benefits and disadvantages of the multimodal biometric authentication system it is suggested to increase testing of the number of attacks and the variety of the spoofing techniques. In the both cases the integrity and confidentiality of system could be reviewed and provide a better understanding of the security and privacy level for multimodal systems. Moreover, some additional security improvement layers such as liveliness detection may be added to the biometric matching process and described the dynamics of system response under attacks with an additional module.

Appendix A

System Architecture

Appendix A describes the architecture of the constructed biometric authentication model, developed and tested using Speaking Faces and the noted validation datasets (Ch. 3). The first two sections present the pseudo-code of the face module with additional required setups and all included libraries and internal functionality. The last section describes the pseudo-code of the voice module architecture with all details. The system was written in the Python language on the Google Colab service. The detailed architecture of unimodal components can be found in Ch.4.

A.1 Face Module *(AI)*



Unimodal biometric component for face recognition by Zhang et al. [59]

Fig.A.1 represents the original architecture of the model provided by Zhang et

al.[59]. The recreated model that was used in the thesis work has the same architecture, excluding a preprocessing stage; please refer to Ch.5.

The list of libraries and additional components that were used: numpy, cv2, cv2_imshow component from google.colab.patches library, os, itemgetter from operator library and time libraries.

The Face module consist of 4 functions: LBP pixel calculation, distance calculation, creation of database and testing database.

A.1.1 LBP Pixel Calculation

This function takes three arguments: image and x and y position of the pixel and return new value of the pixel.

```
def lbp_calculated_pixel(img, x, y):
    center = img[x][y]
    H,W=img.shape
    val_ar = []
    # check all neighbors of the pixel
    for i in range (-1,2):
      for j in range (-1,2):
        if i==0 and j==0: # center pixel
          val_ar.append(0)
        else:
        #pixels who outside of the image boundaries
          if x+i==H or y+j==W or x+i<0 or y+j<0:
             val_ar.append(0)
          elif img[x+i,y+j]<center:
            val_ar.append(0)
          else:
            val_ar.append(1)
```

```
# convert binary values to decimal
power_val = [128, 64, 32, 2, 0, 1, 16, 8, 4]
val = 0
for i in range(len(val_ar)):
    val += val_ar[i] * power_val[i]
return val
```

Fig. A-1 represents the coding scheme of LBP recalculation for the value in power_val array that multiplied to neighbor pixels value which represents as 0 for those who are less intensive than center pixel and 1 for those who have same intensity or greater from center pixel.

128	64	32
2	0	1
16	8	4

Figure A-1: LBP coding principle for pixel calculation

A.1.2 Distance Calculation

For distance calculation the following metrics were used: Euclidean and city-block or Manhattan distances. The calculation requires two arguments comprised of an image from the training database and the current image for recognition, and returns the distance score between the two images. First, the images are normalized by dividing all pixels values by 255 and then reshaped into 250*250 images with an inter-are interpolation method by using the cv2.reshape component. The function calculates Euclidean and city-block metrics by calculating the square of difference and absolute value of difference between pixels within the same position respectively and adds this value to the total distance value.

A.1.3 Creation of Database

This function creates an entire database for the storage of facial features of the authenticated users and adds the acquired data of new users. The function itself has a path to data storage and buffer storage and checks if the buffer exists. The code tests for each user whether it exists in the buffer, and if not, it creates the buffer.

```
def db_creation():
# Provide path to buffer and data storages
  parent_dir = 'Data storage'
  parent_dir = os.path.abspath(parent_dir)
  buffer_dir= os.path.abspath('Buffer')
  if not os.path.exists(buffer_dir):
    os.makedirs(buffer_dir)
#creates face cascade by passing xml Haar template
  face_cascade = cv2.CascadeClassifier('Haar cascade template')
#creates subject list and check whether subject is in buffer
  sub_id=[]
  for subjects in os.listdir(parent_dir):
    sub_id.append(subjects)
    b_dir=os.path.join(buffer_dir,subjects)
    if not os.path.exists(b_dir):
        os.makedirs(b_dir)
        i_dir=os.path.join(parent_dir,subjects)
    #create feature image for each subject who is not in buffer
        for images in os.listdir(i_dir):
            img_dir=os.path.join(ifront_dir,images)
            img = cv2.imread(img_dir,0)
            #Face detection on the image
            faces = face_cascade_frontal.detectMultiScale(img,
                                                  scale,neigh)
```

```
img_bgr=img[faces[0][1]:faces[0][1]+faces[0][3],
    faces[0][0]:faces[0][0]+faces[0][2]]
height, width= img_bgr.shape
#Creation feature image
img_gray = img_bgr
img_lbp = np.zeros((height, width), np.uint8)
for i in range(0, height):
  for j in range(0, width):
    img_lbp[i, j] = lbp_calculated_pixel(img_gray, i, j)
#saving feature image in buffer
newimg_dir=os.path.join(b_dir,str(images))
cv2.imwrite(newimg_dir, img_lbp)
```

A.1.4 Testing System

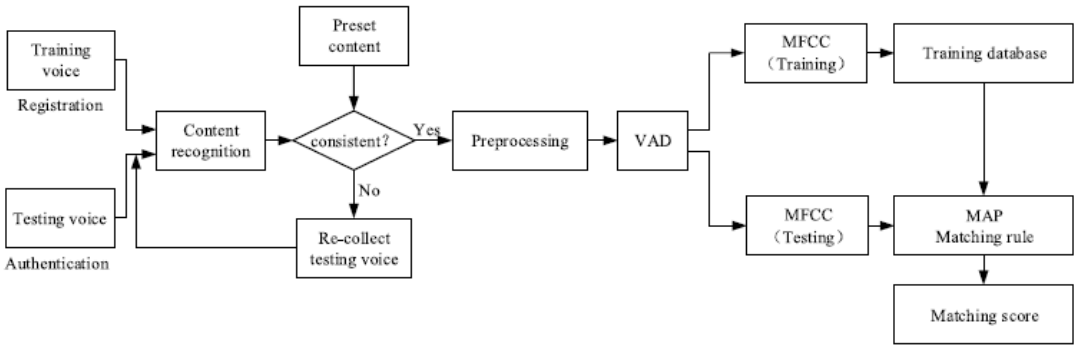
The testing system is similar to database creation. The obtained image went through the face detection and feature extraction process, as it can be reviewed from the database creation code. However, instead of saving the image in the buffer, it is compared with all images in the buffer within the same position and returns a nested list which keeps the result of each compared user and the distance between the inserted image and the authorized user in the database.

A.1.5 Additional Parameters for Georgia Tech Face Database

The major difference between models created for SpeakingFaces and Georgia Tech Face datasets is the inclusion of an additional preprocessing stage for the GTF images. First, the image is loaded in the RGB scale using the imread function. Then the image is manually converted to the gray scale and proceeds with Histogram equalization and picture normalization. These procedures were implemented to reduce the effect of the background noise in the picture during the face detection step. Due to the preprocessing of the image, the minimal face feature image size was set to 140*140.

```
img = cv2.imread(img_dir)
gray = cv2.cvtColor(img, cv2.COLOR_RGB2GRAY)
equ = cv2.equalizeHist(gray)
final_img = cv2.normalize(equ,  equ, 0, 255, cv2.NORM_MINMAX)
faces = face_cascade_frontal.detectMultiScale(final_img,scale,
        neigh, minSize=(140,140))
```

A.2 Voice Module *(AK)*



Unimodal biometric component for voice identification by Zhang et al. [59]

To implement the part where the user is recognized by speech, the Python wave and soundfile audio file reading libraries were used. At the preprocessing stage, the pywt and librosa libraries were imported to decompose the file into wavelets and to separate empty frames from those carrying a signal respectively. Further, the features were extracted using the sklearn and python_speech_features libraries, and the user recognition model also used the GMM implementation from sklearn.

The entire architecture of the speech module code consists of data collection functions, DWT, VAD, MFCC features extraction, training and testing of the model, as well as its validation on the TIMIT dataset.

A.2.1 Collection of Data

To read the digital signal from the audio files, separate training and test text files were created. All samples from the first trial of the dataset assembly are written to the file for training the model and 20% of the test files in the amount of 12 voice samples are written to the second file. For the convenience of tracking the user's ID at the training stage and model test, all files were sorted in alphabetical order.

```
def get_file_names(file, mode):
    basepath = audio_dir
    fw = open(voice_module_dir + file, "w")
    for d in os.listdir(basepath):
        if d != 'intruders':
            count = 0
            if mode == 'train':
                for f in os.listdir(basepath + d + trial_1_dir):
                    fw.write(f + "\n")
            if mode == 'test':
                for f in os.listdir(basepath + d + trial_2_dir):
                    if count >= 12:
                        break
                    fw.write(f + "\n")
                    count += 1
    fw.close()
    sorting(file)
```

A.2.2 DWT Application

In this function, files with an increased level of noise are cleaned by decomposition into separate wavelets. The signal and its sample rate are read into separate variables. For wavelet creation the decomposition level of the wavelet is determined along with

the noise level threshold. Then the signal is decomposed into coefficients, which are subsequently passed through the threshold. On the last lines of the code, the signal is reconstructed with new data without noise and is rewritten into a file.

```
def DWT(source, path, dest):
    sample_rate, data = read(source + path)
    t = np.arange(len(data))/float(sample_rate)
    data = data/max(data)

    w = pywt.Wavelet('sym4')
    maxlev = pywt.dwt_max_level(len(data), w.dec_len)
    threshold = 0.2

    coeffs = pywt.wavedec(data, 'sym4', level=maxlev)
    for i in range(1, len(coeffs)):
        coeffs[i] = pywt.threshold(coeffs[i], threshold*max(coeffs[i]))

    datarec = pywt.waverec(coeffs, 'sym4')
    wf.write(dest + path, sample_rate, datarec)
```

A.2.3 VAD Application

This function separates empty frames from those who carry any signal frequencies using the librosa library. The file information was recorded in the variables date and sample rate. From the selected main signal data, intervals carrying speech amplitudes are determined and recorded in a separate array. Using this array, the signal is rewritten to the original file.

```
def VAD(source, path, dest):
    data, rate = librosa.load(source + path)
```

```
intervals = librosa.effects.split(data, top_db=20,
                                  frame_length=2048, hop_length=2)

no_silence = []

for interval in intervals:
        no_silence.append(data[interval[0]:interval[1]])

no_silence = np.asarray(no_silence)

file_vad = np.concatenate(no_silence)
sf.write(dest + path, file_vad, rate)
```

A.2.4 MFCC Feature Extraction

In this part, the function of calculating delta values is used after collecting the MFCC characteristics from a separate implemented code. The necessary features are read from the audio file and then scaled in the desired interval. Then, using a coded formula, delta values are calculated and written into one stack along with the features themselves. As a result, each stack of all audio files is transferred to the model for further user recognition.

A.2.5 Training and Testing Parts

During the training stage, the features of each user file are collected into one array of features, on the basis of which a separate model is built. Based on the user number, the configured model is written to a file with the corresponding name in the folder with all models on Google Drive. The feature array is reset to write data for the next user. Then, during system testing, each file from the test sample is compared with each available model, where the likelihood score is calculated. The array of likelihood values is sorted from largest to smallest, which thus serves to rank-

order the candidates. To calculate the prediction accuracy among all user files, the percentage of finding the correct target among the top 5 of all similarity values and finding the required target in the first place are calculated separately. All predictions and similarity values are stored in the corresponding arrays.

A.3 Fusion Score

The last step of the authentication system is the calculation of the fusion score. Figure A-2 demonstrated which input required in the testing module for the fusion score and which output is produced by system. From unimodal components two arrays of the all users authorized in the system and corresponding similarity scores were provided for the fusion: distance score or f_{score} from face module and likelihood score or v_{score} from voice module.

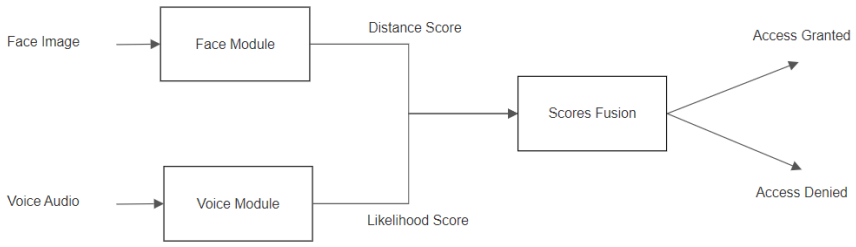


Figure A-2: Flowchart of the fusion score calculation process

In the scope of this thesis work two fusion scores were implemented: ranking based and threshold based.

A.3.1 Threshold Based

Threshold based fusion score was created based on description provided by Zhang et al.[59]. The function takes as input f_{score}, v_{score} arrays and username that authorized in the database. Both f_{score} and v_{score} pass through min-max score normalization process and then the corresponding scores for provided username are selected. The

following calculations are used in the threshold based fusion function:

$$t_1 = a * v\_score + (1 - a) * f\_score,$$

$$t_2 = (a * v\_score) * ((1 - a) * f\_score),$$

$$f(t_1) = \begin{cases} 1, t_1 \geq t_{sum} \\ 0, t_1 \leq t_{sum} \end{cases}$$

$$f(t_2) = \begin{cases} 1, t_2 \geq t_{pro} \\ 0, t_2 \leq t_{pro} \end{cases}$$

The values of t_1 and t_2 are compared with thresholds which was defined in Chapter 4 with parameter a =0.5 for equal treating voice and face component results. The final decision score is evaluated as:

$$f_{decision} = f(t_1) * f(t_2),$$

where $f_{decision}$ returns 0 means access denied and for $f_{decision}$ equals to 1 means access granted.

A.3.2 Ranking Based

Ranking based fusion score requires only the sorted usernames of both f_{score} and v_{score} by similarity score to the inputted biometric data. Top N ranking system extracts first closest N usernames from the f_{score} and v_{score} and compares whether both set of usernames share the common user. If no common user was identified, system denied the access for submitted biometric data.

If there are two or more common users from N closest by face and voice data, the sum of indexes are compared to that set. User with the lowest sum is returned to the system by displaying that access provided to the returned user. For the draw sum of indexes situation, where two or more users have the same sum, the preference is given for user which voice index is smaller. The ranking based fusion score is a biased to the voice component of the system due to the higher accuracy provided by unimodal

voice component.

Bibliography

[1] Madina Abdrakhmanova, Askat Kuzdeuov, Sheikh Jarju, Yerbolat Khassanov, Michael Lewis, and Huseyin Atakan Varol. Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. *arXiv preprint arXiv:2012.02961*, 2020.

[2] Anter Abozaid, Ayman Haggag, Hany Kasban, and Mostafa Eltokhy. Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. *Multimedia Tools and Applications*, 78(12):16345–16361, 2019.

[3] Mohammed Abuhamad, Ahmed Abusnaina, DaeHun Nyang, and David Mohaisen. Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey. *IEEE Internet of Things Journal*, 8(1):65–84, 2020.

[4] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers. Speaker identification evaluation based on the speech biometric and i-vector model using the timit and ntimit databases. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2017.

[5] Soad Almabdy and Lamiaa Elrefaei. Deep convolutional neural network-based approaches for face recognition. *Applied Sciences*, 9(20):4397, 2019.

[6] Sara Amini, Vahid Noroozi, Amit Pande, Satyajit Gupte, Philip S Yu, and Chris Kanich. Deepauth: A framework for continuous user re-authentication in mobile apps. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2027–2035, 2018.

[7] Hagai Aronowitz, Min Li, Orith Toledo-Ronen, Sivan Harary, Amir Geva, Shay Ben-David, Asaf Rendel, Ron Hoory, Nalini Ratha, Sharath Pankanti, et al. Multi-modal biometrics for mobile authentication. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[8] B. Ayotte, M. Banavar, D. Hou, and S. Schuckers. Fast free-text authentication via instance-based keystroke dynamics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):377–387, 2020.

[9] Kadir Sercan Bayram and Bülent Bolat. Multibiometric identification by using ear, face, and thermal face. *EURASIP Journal on Image and Video Processing*, 2018(1):1–8, 2018.

[10] Kemal Bicakci, Oguzhan Salman, Yusuf Uzunay, and Mehmet Tan. Analysis and evaluation of keystroke dynamics as a feature of contextual authentication. In *2020 International Conference on Information Security and Cryptology (ISC-TURKEY)*, pages 11–17. IEEE, 2020.

[11] Andrew Boles and Paul Rad. Voice biometrics: Deep learning-based voiceprint authentication system. In *2017 12th System of Systems Engineering Conference (SoSE)*, pages 1–6. IEEE, 2017.

[12] Neel Ramakant Borkar and Sonia Kuwelkar. Real-time implementation of face recognition system. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pages 249–255. IEEE, 2017.

[13] Anurag Chowdhury, Yousef Atoum, Luan Tran, Xiaoming Liu, and Arun Ross. Msu-avis dataset: Fusing face and voice modalities for biometric recognition in indoor surveillance videos. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3567–3573. IEEE, 2018.

[14] A. Dustor. Speaker verification with timit corpus - some remarks on classical methods. In *2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 174–179, 2020.

[15] Muhammad Ehatisham-ul Haq, Muhammad Awais Azam, Usman Naeem, Yasar Amin, and Jonathan Loo. Continuous authentication of smartphone users based on activity pattern recognition using passive mobile sensing. *Journal of Network and Computer Applications*, 109:24–35, 2018.

[16] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel. On the vulnerability of speaker verification to realistic voice spoofing. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2015.

[17] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[18] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, and D. Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 01 1993.

[19] D Harikrishnan, N Sunil Kumar, R Shelbi Joseph, Kishor Nair, R Nishanth, and Abin John Joseph. Fpga implementation of fast & secure fingerprint authentication using trsg (true random and timestamp generator). *Microprocessors and Microsystems*, 82:103858, 2021.

[20] Janson Hendryli and Dyah Erny Herwindiati. Voice authentication model for one-time password using deep learning models. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology*, pages 35–39, 2020.

[21] Kouki Hongo and Hironobu Takano. Personal authentication with an iris image captured under visible-light condition. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2266–2270. IEEE, 2018.

[22] Mohammad Alamgir Hossain and Basem Assiri. Emotion specific human face authentication based on infrared thermal image. In *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pages 1–6. IEEE, 2020.

[23] Luo Jiang, Juyong Zhang, and Bailin Deng. Robust rgb-d face recognition using attribute-aware loss. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2552–2566, 2019.

[24] Rick Joyce and Gopal Gupta. Identity authentication based on keystroke latencies. *Communications of the ACM*, 33(2):168–176, 1990.

[25] Himanka Kalita, Emanuele Maiorana, and Patrizio Campisi. Keystroke dynamics for biometric recognition in handheld devices. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, pages 410–416. IEEE, 2020.

[26] P Kasemsumran, S Auephanwiriyakul, and N Theera-Umpon. Face recognition using string grammar nearest neighbor technique. *Journal of Image and Graphics*, 3(1):6–10, 2015.

[27] Atul N Kataria, Dipak M Adhyaru, Ankit K Sharma, and Tanish H Zaveri. A survey of automated biometric authentication techniques. In *2013 Nirma university international conference on engineering (NUiCONE)*, pages 1–6. IEEE, 2013.

[28] Sukhchain Kaur and Reecha Sharma. An intelligent approach for anti-spoofing in a multimodal biometric system. *Int. J. Comput. Sci. Eng*, 9:522–529, 2017.

[29] Mate Krišto and Marina Ivasic-Kos. An overview of thermal face recognition methods. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1098–1103. IEEE, 2018.

[30] Yangyang Lian, Zhihui Wang, Hanqing Yuan, Lifang Gao, Zhuozhi Yu, Wenwei Chen, Yifei Xing, Siya Xu, and Lei Feng. Partial occlusion face recognition method based on acupoints locating through infrared thermal imaging. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 1394–1399. IEEE, 2020.

[31] Ze Lu, Xudong Jiang, and Alex Kot. A novel lbp-based color descriptor for face recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1857–1861. IEEE, 2017.

[32] Chao Ma, Ngo Thanh Trung, Hideaki Uchiyama, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Adapting local features for face detection in thermal image. *Sensors*, 17(12):2741, 2017.

[33] 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.

[34] Fabian Monrose and Aviel D Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation computer systems*, 16(4):351–359, 2000.

[35] Ramya T. N and Veena M B. Analysis of polynomial co-efficient based authentication for 3d fingerprints. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–6, 2020.

[36] A.V Nefian, M Khosravi, and MH Hayes. Real-time human face detection from uncontrolled environments. *SPIE Visual Communications on Image Processing*, 1997.

[37] Mohammad Ali Nematollahi, Hamurabi Gamboa-Rosales, Francisco J Martinez-Ruiz, I Jose, Syed Abdul Rahman Al-Haddad, and Mansour Esmaeilpour. Multifactor authentication model based on multipurpose speech watermarking and online speaker recognition. *Multimedia Tools and Applications*, 76(5):7251–7281, 2017.

[38] JAC Nunes, FP Ferreira, and TBA de Carvalho. Waveletfaces and linear regression classification for face recognition. In *2017 Workshop of Computer Vision (WVC)*, pages 144–149. IEEE, 2017.

[39] 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.

[40] Xiujie Qu, Tianbo Wei, Cheng Peng, and Peng Du. A fast face recognition system based on deep learning. In *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 289–292. IEEE, 2018.

[41] FA Rezaur rahman Chowdhury, Quan Wang, Ignacio Lopez Moreno, and Li Wan. Attention-based models for text-dependent speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5359–5363. IEEE, 2018.

[42] Ricardo N Rodrigues, Niranjan Kamat, and Venu Govindaraju. Evaluation of biometric spoofing in a multimodal system. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–5. IEEE, 2010.

[43] Zhang Rui and Zheng Yan. A survey on biometric authentication: Toward secure and privacy-preserving identification. *IEEE Access*, 7:5994–6009, 2018.

[44] Antu Saha, Joydev Saha, and Barshon Sen. An expert multi-modal person authentication system based on feature level fusion of iris and retina recognition. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–5. IEEE, 2019.

[45] Bayu Aji Sahar, Azel Fayyad Rahardian, Elvayandri Muchtar, et al. Fingershield atm–atm security system using fingerprint authentication. In *2018 International Symposium on Electronics and Smart Devices (ISESD)*, pages 1–6. IEEE, 2018.

[46] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Marta Gomez-Barrero, Raymond NJ Veldhuis, Luuk Spreeuwers, Maikel Schils, Davide Maltoni, Patrick Grother, Sebastien Marcel, et al. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2017.

[47] Anna Sidorova and Konstantin Kogos. Voice authentication based on the russian-language dataset, mfcc method and the anomaly detection algorithm. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 537–540. IEEE, 2020.

[48] Shivshakti Singh, Aditi Inamdar, Aishwarya Kore, and Aprupa Pawar. Analysis of algorithms for user authentication using keystroke dynamics. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 0337–0341. IEEE, 2020.

[49] E Sujatha and A Chilambuchelvan Nil. Multimodal biometric authentication algorithm at score level fusion using hybrid optimization. *Wireless Communication Technology*, 2(1):1–12, 2018.

[50] Shahad Sultan and Mayada Faris Ghanim. Human retina based identification system using gabor filters and gda technique. *Journal of Communications Software and Systems*, 16(3):243–253, 2020.

[51] HD Supreetha Gowda, G Hemantha Kumar, and Mohammad Imran. Multimodal biometric system on various levels of fusion using lpq features. *Journal of Information and Optimization Sciences*, 39(1):169–181, 2018.

[52] Florentin Thullier, Bruno Bouchard, and Bob-Antoine J Menelas. A text-independent speaker authentication system for mobile devices. *cryptography*, 1(3):16, 2017.

[53] Dimitra Triantafyllou, Georgios Stavropoulos, and Dimitrios Tzovaras. Iris authentication utilizing co-occurrence matrices and textile features. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE, 2019.

[54] Ka-Wing Tse and Kevin Hung. User behavioral biometrics identification on mobile platform using multimodal fusion of keystroke and swipe dynamics and recurrent neural network. In *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 262–267. IEEE, 2020.

[55] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.

[56] Dana Weitzner, David Mendlovic, and Raja Giryes. Face authentication from grayscale coded light field. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2611–2615. IEEE, 2020.

[57] Heng Zhang, Vishal M Patel, and Rama Chellappa. Low-rank and joint sparse representations for multi-modal recognition. *IEEE Transactions on Image Processing*, 26(10):4741–4752, 2017.

[58] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1080–1091, 2016.

[59] Xinman Zhang, Dongxu Cheng, Pukun Jia, Yixuan Dai, and Xuebin Xu. An efficient android-based multimodal biometric authentication system with face and voice. *IEEE Access*, 8:102757–102772, 2020.

[60] Xinman Zhang, Kunlei Jing, Yixuan Dai, and Xuebin Xu. Face biometric identity authentication system. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 1473–1477. IEEE, 2018.

[61] Xinman Zhang, Qi Xiong, Yixuan Dai, and Xuebin Xu. Voice biometric identity authentication system based on android smart phone. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 1440–1444. IEEE, 2018.

[62] Yongtuo Zhang, Wen Hu, Weitao Xu, Chun Tung Chou, and Jiankun Hu. Continuous authentication using eye movement response of implicit visual stimuli. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22, 2018.