BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

TESTING THE ROBUSTNESS OF A VOICE BIOMETRICS SYSTEM AGAINST DEEPFAKES

TESTOVÁNÍ ODOLNOSTI SYSTÉMU HLASOVÉ BIOMETRIE VŮČI DEEPFAKES

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. JAKUB REŠ**
AUTOR PRÁCE

SUPERVISOR **Mgr. KAMIL MALINKA, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

144004

Institut:	Department of Intelligent Systems (UITS)
Student:	**Reš Jakub, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Cybersecurity
Title:	**Testing the Robustness of a Voice Biometrics System against Deepfakes**
Category:	Security
Academic year:	2022/23

Assignment:

1. Learn about deepfakes and voice biometrics systems, including relevant ISO standards.
2. Learn the methodologies for testing the robustness of biometric authentication. Focus on voice biometric systems and methods to defend against spoofing (using deepfakes).
3. Propose an appropriate method for testing the robustness of the existing voice biometric system in order to discover system weaknesses when using deepfakes. For testing, use publicly available voice deepfakes datasets.
4. Based on the design, test the selected system. Evaluate the results and their impact on the security of the voice biometrics system. Based on the results, define the weak points of the tested system and discuss possible methods to protect these points.
5. Transform the proposed testing method into a methodology that can be used by authentication system vendors to repeatably test the robustness of voice-based biometric authentication systems against deepfakes.

Literature:

Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa
Singh, “Facial retouching and alteration detection,” in Handbook of Digital Face Manipulation and Detection, pp. 367–
387. Springer, 2022
FIRC Anton a MALINKA Kamil. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: Brno: Association for Computing Machinery, 2022

Requirements for the semestral defence:
Items 1 to 3.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Malinka Kamil, Mgr., Ph.D.**
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2022
Submission deadline:	17.5.2023
Approval date:	3.11.2022

Abstract

Topic of this paper is a methodology of testing the robustness of a voice biometric system against deepfakes. The main problem currently lies in insufficient coverage of testing against the presentation attack using deepfakes in ISO/IEC standards. The aim of this thesis is to cover the hole, resulting from emergence of deepfake technology, by proposing an extended methodology, based on the existing one, that focuses on fixing the issue. The solution of proposed problem started by studying the state of the art for deepfakes and standard practices of biometric system testing. Second, I proposed and documented a method of testing the voice biometric system. The test was designed as a scenario, where the Phonexia voice biometric system is used as a remote verification tool for the voice-as-a-password use-case. For the purpose of demonstration, the online publicly available dataset was used. On top of test design, I set a non-standard metric for the test evaluation to show possibilities of focus on different kinds of deepfakes. After carrying out tests and evaluating results, I formulated the procedure into a generic repeatable methodology, containing practices and recommendations. The contribution of this work lies in incorporating deepfakes into the existing standard methodologies of testing a biometric systems, hence forming and demonstrating a repeatable methodology.

Abstrakt

Tématem této práce je vytvoření metodologie testování odolnosti hlasového biometrického systému vůči deepfakům. Hlavní problém v současné době leží v nedostatečném pokrytí testování proti prezentačním útokům užitím deepfaků ve standardech ISO/IEC. Cílem práce je vyplnění této mezery, vzniklé příchodem technologie deepfaků, navržením metodologie, založené na současných postupech, která se soustředí na pokrytí této problematiky. Řešení navrženého problému začíná studií nejnovějšího stavu oblasti deepfaků a standardních postupů pro testování biometrických systémů. Druhým krokem je navržení a zdokumentování metody testování hlasového biometrického systému. Test byl navržen jako scénář, ve kterém je hlasový biometrický systém Phonexia použit jako nástroj pro vzdálenou verifikaci použitý pro hlas-jako-heslo. Pro účely demonstrace byl použit veřejně online dostupná datová sada. Mimo samotný návrh testu jsem také zavedl nestandardní metriku vyhodnocení pro ukázku možností zaměření na různé typy deepfaků. Po provedení a vyhodnocení testů jsem zformuloval postup do obecné opakovatelné metodologie, obsahující praktiky a doporučení. Přínos této práce leží v zapracování deepfaků do existujících standardních metodologií testování biometrických systémů a tak formování a demonstrování opakovatelné metodologie.

Keywords

deepfake, methodology, testing, spoofing, biometic system

Klíčová slova

deepfake, metodika, testování, spoofing, biometický systém

Reference

REŠ, Jakub. *Testing the robustness of a voice biometrics system against deepfakes.* Brno, 2023. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.

Rozšířený abstrakt

Deepfakes jsou nově rostoucí problém. Mimo všech možností masové manipulace či vydírání, jsou deepfaky často také využívané jako nástroj pro vydávání se za jiné osoby a následné útoky na biometrické systémy – spoofing útoky. Z tohoto důvodu jsou vývojáři nuceni jednat rychle a přizpůsobit své existující systémy na obranu proti novým útokům. Cyklus vývoje se skládá nejen z implementace nových detekčních metod, ale také obsahuje fázi testování. Testování je běžně prováděno dle standardních metodologií. Současné standardy jsou ovšem často zastaralé a nezvažují deepfaky jako možnost zdroje útoku. Tato práce cílí na nápravu tohoto nedostatku návrhem metodologie testování proti deepfakům.

Nyní standardy pro testování biometrických systémů obsahují ověřené obecné postupy, kterými by se měly firmy řídit při hodnocení jejich produktů. Tyto standardy ale bohužel zvažují pouze konvenční metody tvorby falzifikátů. Stejně tak jako běžné metody je třeba brát zřetel i na deepfaky a věnovat jim dostatek pozornosti. Tato práce se soustředí na demonstraci návrhu metody testování and její provedení pro dodaný biometrický systém, následováno zobecněním dosavadního postupu v metodiky pro opakovatelné testování odolnosti hlasových biometrických systémů vůči spoofing útoku pomocí deepfaků založené na existujících standardech.

Jak bylo zmíněno, existující řešení, primárně ve formě standardů, obsahují obecné postupy testování biometrických systémů. Standardy, které tato práce využívá, jsou primárně ISO/IEC 19795-1:2006 [16] and ISO/IEC 19795-2:2007 [17].

Z hlediska testování specificky deepfaků, existují mnohé články o testování detekčních metod, ale velmi málo, až žádné, které se věnují testování implementovaných metod jako součást systému. I přes to nabízejí užitečné informace a postupy, které stojí za inspiraci.

Řešení navržené touto prací spočívá v kombinaci existujících metod testování, jak jsou navrženy ve standardech, s přístupy k testování metod detekce deepfaků. Využití dlouhodobě zavedených a ověřených praktik a jejich rozšíření o užitečné dodatky týkající se deepfaků a jejich hodnocení se zdá být jako správný přístup.

Za účelem formulace metodologie a demonstrace postupu jsou navrhl vlastní metodu testování dodaného, komerčně používaného systému hlasové biometrie Phonexia. Metoda je založena na postupech doporučovaných standardy s doplňujícím zaměřením na použití veřejně online dostupných datových sad deepfaků a navržení nestandardní metriky jako příklad možností sledování vlivu různých typů deepfaku dle metody jejich tvorby.

Oblasti navržené metody jsou následující:

- Cíl – cíl testování (vyhodnocení odolnosti biometrického systémů vůči různým skupinám deepfaků s cílem stanovení odolnosti vzhledem k metodě jejich tvorby)
- Případ použití – případ použití testovaného systému (verifikace, hlas-jako-heslo)
- Model útočníka – motiv, příležitost a prostředky potencionálního útočníka
- Scénář – shrnutí předešlých bodů do testovacího scénáře
- Datová sada – veřejně dostupná datová sada, která byla použita (ASVspoof19)
- Metriky – metriky hodnocení biometrického systému (standardní/vlastní)

V souladu s navrženou metodou bylo provedeno testování systému jako nástroje pro vzdálenou verifikaci použitého pro hlas-jako-heslo. Práce popisuje postup, dle kterého byl test proveden, včetně popisu prostředí, vlastností testovaného systému a popisu komunikace

mezi testovacím nástrojem a testovaným subjektem. Dále práce popisuje experimenty, jak byl simulován navržený scénář, jak byla modifikována data ze zvolené datové sady tak, aby splňovala požadavky systému a jaké byly zaznamenané výsledky. Nakonec je popsáno vyhodnocení výsledků a přínos navržených metrik k identifikaci možných slabých míst.

Výsledná osnova navrhované metodologie se skládá z pěti hlavních částí:

- Fáze plánování – první fáze každého testování. Tato fáze se zaměřuje na sběr důležitých informací of systému, potenciálních útočníků a definici hlavního cíle testování.
- Sběr dat – tato fáze je o doporučeních týkajících se shánění správné datové sady. Ať už se jedná o sběr vlastních dat nebo použití existujících.
- Provedení testů – tato fáze je typicky stejná jako ji navrhují standardy.
- Vyhodnocení – fáze o metrikách a zdůvodnění použití vlastních.
- Interpretace – poslední krátká fáze o relevanci vyhodnocených výsledků dle statistických pravidel daných standardy.

Testing the robustness of a voice biometrics system against deepfakes

Declaration

I hereby declare that this Master's thesis was prepared as an original work by the author under the supervision of Mr. Mgr. Kamil Malinka Ph.D. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

.....
Jakub Reš
May 11, 2023

Acknowledgements

I want to thank my supervisor Mgr. Kamil Malinka Ph.D. for his help, motivation, valuable advice and support during the writing of this thesis.

I would also like to thank the group of people at Phonexia for their helpfulness and willingness to provide valuable consultation on their system.

And a big thank you to my family and friends for their unending moral support.

Contents

Chapter 1

Introduction

Deepfakes are an emerging problem. Besides all the possibilities of mass manipulation or blackmail, deepfakes are also often used as a tool for impersonation and subsequent attacks on biometric systems – spoofing attacks. For this reason, developers are forced to act quickly and adapt their existing systems to defend against new attacks. The development cycle not only consists of implementing new detection methods, but also includes a testing phase. Testing is normally performed according to standard methodologies. However, current standards are often outdated and do not consider deepfakes as a possible source of attack. This work aims to address this shortcoming by proposing a methodology for testing against deepfakes.

Standards for testing biometric systems now contain well-established general practices that companies should follow when evaluating their products. Unfortunately, however, these standards only consider conventional methods of creating forgeries. As well as conventional methods, deepfakes should also be considered and given sufficient attention. This paper focuses on demonstrating the design of a testing method and its implementation for a delivered biometric system, followed by a generalization of the existing approach into a methodology for repeatable testing of the robustness of voice biometric systems to spoofing attack using deepfakes based on existing standards.

As mentioned, existing solutions, primarily in the form of standards, contain general procedures for testing biometric systems. The standards used in this work are primarily ISO/IEC 19795-1:2006 [16] and ISO/IEC 19795-2:2007 [17].

In terms of testing specifically deepfakes, there are many papers on testing detection methods, but very few, if any, that address testing implemented methods as part of a system. Despite this, they offer useful information and techniques that are worth getting inspired by.

The solution proposed in this paper is to combine existing testing methods as proposed in standards with approaches to testing deepfake detection methods. Using long established and proven practices and extending them with useful additions related to deepfakes and their evaluation seems like the right approach.

In order to formulate a methodology and demonstrate the procedure, a custom method is proposed for testing the provided, commercially used Phonexia voice biometrics system. The method is based on standards-recommended procedures, with an additional focus on using publicly available online deepfake datasets and proposing non-standard metrics as an example of the possibilities of monitoring the impact of different types of deepfakes according to the method of their creation.

The areas of the proposed method are as follows:

- Goal – testing goal (to evaluate the robustness of biometric systems to different groups of deepfakes in order to determine the robustness with respect to the method of their creation)
- Use-case – use case of the system under test (verification, voice-as-password)
- Attacker model – motive, opportunity and means of a potential attacker
- Scenario – summary of the previous points into a test scenario
- Dataset – publicly available dataset that was used (ASVspoof19)
- Metrics – Biometric system evaluation metrics (standard/custom)

In accordance with the proposed method, the system was tested as a remote verification tool used as a voice-as-password. The paper describes the procedure followed to perform the test, including a description of the environment, the properties of the system tested, and a description of the communication between the test tool and the test subject. The thesis also describes the experiments, how the proposed scenario was simulated, how the data from the chosen dataset was modified to meet the system requirements, and what the recorded results were. Finally, the evaluation of the results and the contribution of the proposed metrics to the identification of potential vulnerabilities are described.

The final outline of the proposed methodology consists of five main parts:

- Planning phase – the first phase of any testing. This phase focuses on gathering relevant information of the system, potential attackers and defining the main objective of testing.
- Data collection – this phase is about recommendations regarding gathering the right dataset. Whether it is collecting your own data or using existing data.
- Execution of tests – this phase is typically the same as suggested by the standards.
- Evaluation – the phase about metrics and justification for using custom ones.
- Interpretation – the last short phase on the relevance of the results evaluated according to the statistical rules given by the standards.

Chapter 2

Deepfakes

The first section deals with the general topic of deepfake – what is it and what are the consequences of the spread of this technology. This is followed by a section covering the origins and the technologies behind it. The penultimate section looks at the existing datasets available – primarily from the perspective of basic properties. Finally, there is a section on available tools for deepfake voice creation.

2.1 What is deepfake

Fake digital media generated by deep neural network, commonly referred to as *deepfake*, is a subcategory of fake or altered media. The main difference of this technology is the modern way of generating forgeries using deep neural networks in order to achieve results potentially indistinguishable from the original.

2.1.1 Current state

Currently, voice deepfake technology is still far from being fully explored. As for the area of face deepfakes, the scientists are futher, but still not near the end. New methods of both creation and detection are being developed every year. We are in the middle of a race between makers and detectors for who has the upper hand in terms of detection – creating realistic, undetectable forgeries (perhaps a malicious ones) vs. effective detection methods that are able to differentiate between fake and real digital content.

But this technology is becoming an inevitable part of our lives. Today and every day we can encounter a considerable amount of media created in this way in the Internet environment. Many of them are already difficult for people to recognise.

Many social media users enjoy making satirical or funny videos of famous people, politicians or friends. Others, however, see this technology as an opportunity to abuse and manipulate people.

A prime example to showcase the current possibilities is a channel on the social network YouTube run under the name *@unreal_keanu*[1]. The channel's content is directed at short videos depicting an anonymous character with an artificial head, allegedly using deepfake technology, of actor Keanu Reeves.

[1] https://www.youtube.com/@unreal_keanu

While current state of deepfake technology may seem stunning, there are still challenges to overcome. One of them can be seen in the most-right picture of Figure 2.1 – low resolution for complex deepfakes.

Figure 2.1: Deepfakes of Keanu Reeves from channel @unreal_keanu

2.1.2 Why are deepfakes problem

While the deepfake technology may seem like a spectacular advance for digitising many areas of life, it can, and already is, causing considerable harm. There has been cases of people getting blackmailed, having their identity stolen and used with nefarious intentions or being manipulated by forgeries aimed at spreading alarmist messages. The following are examples of areas where deepfakes could cause serious damages.

Politics

Politicians are a popular target of deepfake attacks. The dissemination of fake videos of specific politicians saying false information and immoral, even illegal things can strongly affect the careers and lives of the individuals in question [41]. In addition to defamation and general public outrage against selected politicians, such deepfake videos could also potentially influence presidential elections [25], for example, and thereby threaten the very foundations of democracy.

Another example of the abuse of deepfakes to manipulate masses of soldiers to lay down their arms. The video appeared on social media at the beginning of the Russia-Ukraine war. In it, the attacker displayed a fake image of President Zelensky speaking to Ukrainian soldiers. He abused the influence of the president to manipulate military troops and influence the course of the war [6].

Blackmailing

An activity that initially began as a fun way for individuals to pass the time on social media [12] soon became rightfully feared and very dangerous for ordinary people. Person-swapping within videos, specifically pornographic videos, has become a modern tool for blackmailing individuals [28]. All an attacker needs is a few photographs or a short video of the victim, whose face is then superimposed on the actor/actress using the face-swap technique. Using

material that can be created at almost zero cost they then blackmail their target and try to get as much money out of the victim as possible.

Identity theft

In June, 2022 IC3 issued a report [13] that a new trend in identity theft and subsequent deepfake abuse has emerged. Fraudsters are attempting, using deepfake images of real people on the internet and stolen personal information, to obtain remote work-from-home positions.

The FBI has not stated any clear goal of the attackers to achieve through these scams. But it is assumed, given the job positions targeted by the scammers, that they intend to exploit the acquired reputations of people in the IT industry and gain access to sensitive corporate infrastructures.

Justice

Another critical aspect of deepfake is the potential impact on the delivery of justice in the trial of criminals. A forged voice recording or video can influence the court's decision and shift the blame from the real criminal to an innocent victim [26].

Mass manipulation

As already mentioned, it is now possible to manipulate masses of people in many areas using fake news, whether by defaming famous people or vice versa. One such case was the artificially created video of President Ali Bongo, who suffered a stroke and underwent several operations [9]. Due to the lack of information regarding his health, the public believed that the President was not in good condition. However, the published video showed President Ali Bongo giving a speech to the people. This, along with false information about his health, affected the awareness of a large number of people.

Another possible scenario is, for example, the manipulation of the value of shares on the stock market. Creating a deepfake of prominent, well-known economists or leading figures of world banks or stock exchanges advising people about investments or warning them about value crashes [7].

Phishing and scams

A similar area to manipulation is targeted attacks and fraud. While mass manipulation, as the name suggests, targets large numbers of people with the abuse of a public figure, targeted attacks are more personal. This can often involve situations such as identity theft to manipulate a person close to her, such as co-workers, or even impersonating the government and demanding that actions be taken, from which the attacker typically obtains the victim's finances [7].

With rapidly advancing technology, it is possible to create ever more credible forgeries. Already in 2019, it was mentioned at the RSA security conference [34] that it is possible to exploit deepfake to create human-unrecognizable fake media to conduct a **automated** targeted spear phishing attack.

2.1.3 Possible benefits of deepfakes

Despite all the negatives, deepfakes can offer legitimate uses in many industries. Whether it is entertainment or use in areas where visualization of the effect of external influences on an individual is needed. The following are examples of possible beneficial uses of deepfakes.

Movie industry

Deepfakes have wide applications in the field of movie industry. The possibilities of making films with long-dead actors or re-shooting famous scenes [41] are appealing to producers and consumers alike. Beyond these, deepfakes extend the possibilities of dubbing. In the case of loss of an actor's voice or the need for dubbing in different languages, deepfakes can help not only with dubbing the voice track itself [41], but also with lip-synching. However, there still lies the ethics question of such approach.

Entertainment

For many social media users, deepfakes are an endless world of fun. Every day there are entertaining videos showing celebrities or the creators themselves singing famous songs or dancing popular dances. There are already apps available today for creating similar entertainment, such as Avatarify [2].

Another entertainment industry in which deepfakes can be used is the gaming industry. Game developers could enrich the player experience by using custom dubbing of the player character, personalized helpers, or virtual depictions of familiar real-world characters [41].

Learning

The use of deepfakes in the field of education also offers significant opportunities. Children who cannot attend standard classes for mental or physical health reasons can be helped by personal virtual lessons, in which the teacher's face and voice are replaced by the parent's characteristics using face-swap and voice conversion techniques. Such an approach would help to increase the effectiveness of the teaching [31].

Healthcare

Deepfakes can also be used for medical purposes. One of the many uses is to model the appearance of a patient after plastic surgery or sex reassignment. Another is for Alzheimer's patients. Deepfakes are used to model the faces of young loved ones, which are easier for patients to remember [41].

Apart from these, there is also a possible use for therapeutic purposes. When a close loved one is suddenly lost, a model is created to which patients can express unspoken feelings and say goodbye appropriately [31].

Privacy

Deepfakes can also benefit privacy and anonymity. They can be used to modify significant facial or voice characteristics of a subject, hiding them from human and machine recognition. This feature may be valuable for hiding witnesses who wish to remain anonymous [31].

[2] https://avatarify.ai/

2.2 Technology behind deepfake

This section will summarize the origins of deepfakes, from which era the falsification of media originated and, to this end, the use of machine learning. This is followed by a brief, surface-level description of the general deepfake generator technology.

2.2.1 Origin

As suggested earlier, the term *deepfake* is a compound of the words „deep“ and „fake“, referring to the main idea of the subject – a fake created by deep learning.

The concept of fakes and photo faking is as old as photography itself. Ever since the 19th century, faking has been a hot topic in proper circles. It became fully developed with the arrival of photography in the media.

An important milestone was the work of *Bregler*, *Covell* and *Slaney* [8]. In their work, they present the first way to fully automatically edit a video of a speaker to make the person appear to be speaking arbitrary text. Their work uses machine learning to modify significant points of the face and lips to adjust expression appropriately to the spoken words.

Years of study and popularization of neural networks followed. In 2017, a user with the nickname „deepfake“ appears on the social network Reddit. It is after this user that the technology is named from now on. The named user used face-swap technology, introduced in 2016 by [37], to create pornographic videos with actors who had their faces swapped with celebrities. Many others followed and created many fake videos, such as the famous video featuring then-President Barrack Obama titled *„You Won't Believe What Obama Says In This Video!“* [3]. The above-mentioned video shocked the general public and kicked off research in deepfake detection and prevention.

2.2.2 General deepfake synthesizer architecture

As mentioned several times above, deepfakes are generated by neural networks. There is a specific sub-category of neural networks in the background of this technology – generative. Generative neural networks for creating deepfakes are typically formed using multiple neural networks of different types. The types of neural networks currently used are [27]:

- ED – Encoder/Decoder neural network – These are at least two networks forming an encoder and a decoder. Depending on the structure, these networks typically take care of input summarization or reconstruction. Current deepfakes generators use several of these networks [27].
- CNN – Convolutional neural network – Convolutional networks are a forerunner in image information processing. The individual convolutional network layers are trained hierarchically as filters, the polling layers as dimension reducers.
- GAN – Generative adversarial network – According to Yisroel Mirsky and Wenke Lee [27], GANs consist of two networks: a generator and a discriminator. These two networks outperform each other in learning, the generator trying to fool the discriminator, which in turn detects the generator's output.
- RNN – Recurrent neural network – According to Yisroel Mirsky and Wenke Lee [27], the recurrent network is adapted to handle continuous data. The network keeps

[3] https://www.youtube.com/watch?v=cQ54GDm1eL0

its own internal state with respect to the ongoing data. In the field of deepfakes generation, RNNs are typically used for deepfakes of voice.

2.3 Voice deepfakes

Since this thesis focuses on voice systems, only deepfakes of voice will be described and considered in the following. This section focuses on the two main types of deepfakes of voice: TTS (*text-to-speech*) and VC (*voice conversion*). The section will describe the basic principles of these approaches and their importance.

2.3.1 Text-to-speech

The problem of text-to-speech technology (hereafter referred to as TTS) is not directly related to deepfakes. As such, it has been the subject of study for many years. The first attempts at artificial speech synthesis date back to the 1990s.

According to Taylor [35], TTS is the process of text-to-signal encoding, i.e. the input is text, the output is signal/speech. In the field of deepfakes, TTS is a technique of creating fake speech based on a trained model of the speaker and a textual template. Today's trends in TTS are the so-called Multispeaker TTS synthesizers. The principle of these systems is to decouple the speaker encoder from the general speech model [20]. With this approach, the general speech model can be learned (using a wider corpus of data) and then the appropriate embeddings of the speaker can be attached during synthesis.

Based on the speaker's encoder and the input text, the synthesizer generates a spectrogram to present to the vocoder, which generates the appropriate signal.

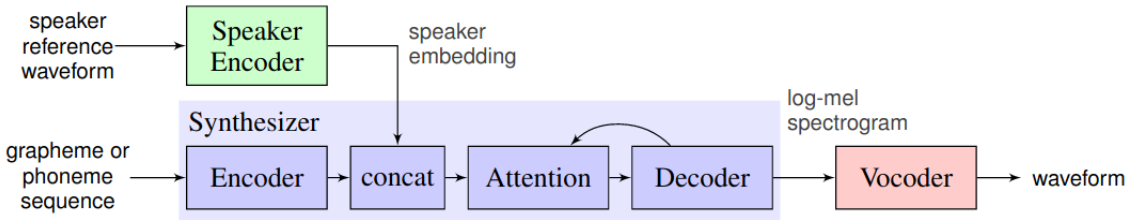


Figure 2.2: General architecture of multi-speaker synthesizer. Each color represents individual component, all trained separately. Image obtained from [20]

2.3.2 Voice conversion

Another category of forgery techniques is a voice conversion. Imitating the voices of other people has been a point of interest for a century now. In 1922, John Q. Stewart wrote an article about the challenge surrounding the voice apparatus synthesis [33].

Voice conversions are created based on the input signal and the characteristics of the target speaker's voice. Thus, the input is not text but speech. The voice conversion technique manipulates only the properties of the voice, not the content of the speech [29].

Voice conversion systems are usually divided into four logical blocks (marked consistently with Figure 2.3):

- (a) Content encoder – a system part that processes input speech and encodes its content

- (b) Style encoder – processes a speech of a different speaker, extracts and encodes its vocal characteristics
- (c) Decoder – decodes the inputs and concatenates them for a future processing
- (d) Vocoder – the final output speech synthesis

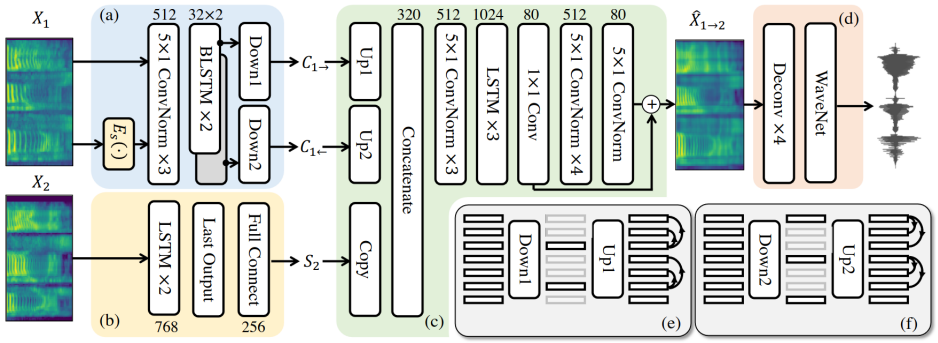


Figure 2.3: Example architecture of a voice conversion system AutoVC. Image retrieved from [30].

2.4 Deepfake datasets

This section lists existing datasets of voice deepfakes. Due to the nature of this thesis, I only focus on a selection of relevant ones from these datasets. The information about each dataset is obtained from the respective papers. These articles will be listed at the beginning of each description.

ASVspoof2019

ASVspoof2019 [38] is the first of two datasets used in this work from the Automatic Speaker Verification and Spoofing Countermeasures challenge. This is a dataset created specifically for the 2019 ASVspoof challenge.

The dataset consists of two main parts depending on the scenario: LA – logical access and PA – physical access. In the case of LA, this is the scenario where the attacker communicates with the given system directly using only the phone (no dedicated sensors or playback device). On the other hand, PA is a scenario where the attacker interacts with the system using a playback device to replay the forgeries to the sensor or telephone, the quality of which is embedded in the samples.

Dataset samples are generated using 17 different TTS and VC systems. According to Todisco et al. [38], the samples are a subset of the VCTK dataset, where they were collected from a total of 107 speakers – 46 male, 61 female. For each scenario in ASVspoof2019, they are subsequently divided into three partitions – *train*, *dev*, and *eval*.

ASVspoof2021

ASVspoof2021 [23] is the second of two datasets used in this work from the Automatic Speaker Verification and Spoofing Countermeasures challenge. It is a dataset created specifically for the ASVspoof challenge in 2021.

Unlike the 2019 version, the dataset now includes a DF – DeepFake section in addition to the standard LA and PA sections. The DF part contains samples created using deepfake technology. This part of the samples is not specifically tailored for the scenario of use with ASV. The DF part of the dataset comes in part from the LA subset of the 2019 dataset, as well as the VCC (Voice Conversion Challenge) for 2018 and 2020.

The evaluation samples of the DF part of the dataset are created by many volunteers and generated using more than 100 algorithms. According to Todisco et al. [38], the samples of the DF part are not intended to deceive the ASV, but only to represent cases where the attacker wants to tarnish the reputation of another person.

FAD

FAD [24] is a Chinese dataset for fake audio detection. The dataset was created in June, 2022 for the purpose of fake audio detection tasks (training/testing) and also for the forensic purposes. To be able to serve in multiple scenarios, the authors made two versions of the samples – clean and noisy. The noisy part of dataset is created adding prepared noises (0dB, 5dB, 10dB, 15dB, 20dB) from the public databases, like PNL 100 Nonspeech Sounds or NOISEX-92.

In total, the dataset contains 431,600 utterances – 215,800 for both clean and noisy parts. Those are divided into four categories – training set (138,400), development set (14400), test set (42000) and unseen test set (21000) with no overlap between training, development and test sets.

Each of the mentions subsets contains real and fake samples. The real samples are collected from OpenSLR [4] and recording their own subjects. The fake part is created using 11 different (representative [24]) methods.

[4] http://www.openslr.org/12/

Chapter 3

Biometric systems

Although biometrics seems like a purely modern technology, its first systematic applications are much older. As early as the 19th century, mankind began systematically collecting biometric data, then primarily fingerprints and handprints, called *Bertillonage*, to identify individuals.

Today, however, we encounter biometrics and biometric systems every day. Authentication to a personal device using a facial image, fingerprint or voice, authentication when entering protected facilities, accessing protected data or banking.

This chapter will describe the voice biometric system and its general architecture. Next the attacks on general biometric system are discussed. This is followed by a description of the general methodology for testing a biometric system. Lastly, due to the scope of this thesis, a description of the voice verification system supplied for the purpose of this thesis – Phonexia.

3.1 Voice biometric systems

A voice biometric system is a system that matches input data, i.e. speech, with a stored template of an individual's voice characteristics. Voice biometrics is considered to be so-called *behavioral* biometrics, that is, it observes how an individual speaks regardless of the speaker's vocabulary[39].

A general biometric system typically supports multiple modes. The two main modes are identification and verification. The identification mode performs a comparison of the input data against its entire database in an attempt to determine the identity of the unknown person (1:N matching). Verification, on the other hand, verifies that the individual is who he or she claims to be, i.e., the system compares the input data against one specific record (1:1 matching).

3.1.1 General architecture

As already mentioned, a biometric system is designed to match input biometric data against stored templates. In order to successfully perform this activity, the biometric verification system needs to contain certain modules to provide the necessary functionality [39].

The first of these modules is the input sample capture module, typically implemented by a set of sensors or cameras. This module is responsible for taking the appropriate sample from the user and sending the data to the next module for feature extraction. The feature extraction module receives the data from the sensors and extracts the relevant properties of

the sample. The features are then passed to the template generation module. This creates a template for comparison/enrollment. The created template is then either stored in the system database in case of enrollment or compared by the template comparison module.

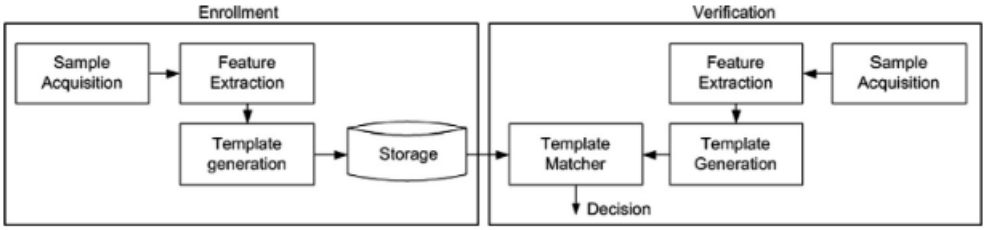


Figure 3.1: General architecture of biometric system. Image retrieved from [39]

3.2 Attacks on biometric system

This section summarizes the types of attacks on the generic biometric system according to each area of vulnerabilities. Next, the presentation attack and its principle is described.

3.2.1 Types of attacks

According to Rubal Jain et al. [19] there are in total eight areas of vulnerabilities of the biometric system. Each of the attacks on these areas are reffered to as *types* of attack. The types of attack are dvivided into two groups based on required knowledge of the target system – direct attacks and indirect attacks.

Direct attacks do not require any specific knowledge of the target system. There is only one type of attack – attack at the sensor.

1. Attack at the sensor – sensor is typically attacked using artificial biometric, image of biometric or damaging the sensor to flood target system with nonsense data

Indirect attacks are the opposite – they do require specific knowledge of how the target system works internally. The rest of the attack types are classified as indirect.

2. Attack at sensor-feature extractor communication – attack involves stealing the transferred biometric data from sensor to feature extractor and replaying them later
3. Attack at feature extractor – attacker convinces the feature extractor to extract specific features instead of features of the presented biometric
4. Attack at feature extractor-matching algorithm communication – same as the type 2 attack, but the attacker steals the extracted features
5. Attack at matching algorithm – attacker convinces the matching algorithm to return high score regardless of the input features
6. Attack at matching algorithm-application communication – attacker modifies the score returned by matching algorithm
7. Attack at matching algorithm-database communication – attacker modifies the content of communication during the template extraction from database

8. Attack at database – attacker inserts custom templates into the database

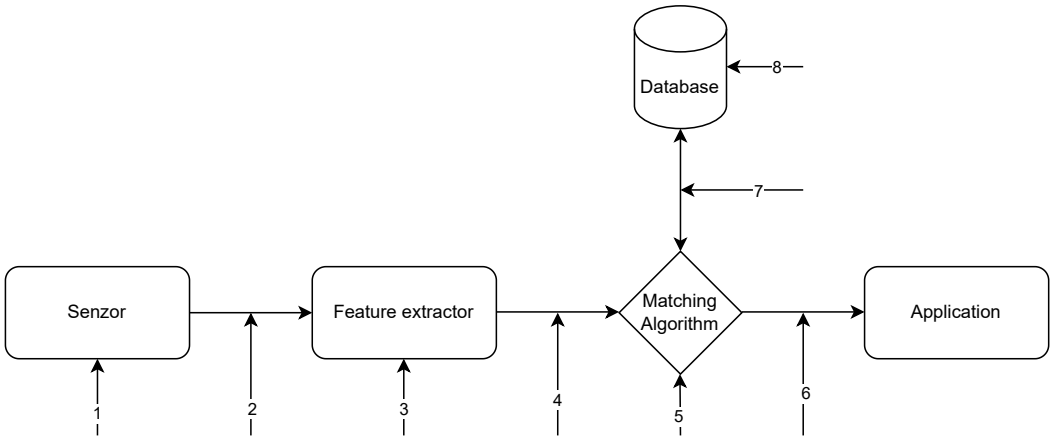


Figure 3.2: Biometric system areas of vulnerabilities. Inspired by [19]

3.2.2 Presentation attack

Presentation attack, often called by the more general name *spoofing* attack, is a type 1 attack on the biometric system – attack at the sensor. The idea behind presentation attack is to present a foreign biometric to the system to fool the system into authenticating the attacker as someone else.

Mostly the attackers use artificial biometrics – a model of hand, fingerprint or even images of these. In the terms of voice, the attacker may use either generators to synthesize a recording or voice conversion system to make themselves sound like victim. Such forgeries are called *spoofs*.

To prevent attackers from attacking using spoofs, developers came with the idea of presentation attack detection systems. A very popular presentation attack detection system is liveness detection. This system tries to detect artificial and inanimate objects by focusing on the signs of people's behaviour or features (for example signs of live in the presented hand or face). The liveness detecting is very problematic when it comes to voice biometric. The usual approach is a conversation between a operator and user in hope of attacker not being prepared for questions and not being able to synthesize the spoofs in real time.

3.3 Testing methodologies

This section describes the general procedure for testing the performance of biometric systems with the added notes on testing the presentation attack detection systems. It describes two basic types of methodology – technology testing vs. scenario testing. It also describes the test design, the general data collection procedure for testing, the measurement process, and finally the output report. The information in this section summarizes the recommendations of the IEC/ISO 19795-1:2006 [16], IEC/ISO 19795-2:2007 [17] and IEC/ISO 30107-3:2017 [18] standards.

3.3.1 Types of methodologies

In general, we can distinguish two general types of methodologies for testing the system's performance – technology testing and scenario testing. The basic distinction can be summarized as follows.

Technology evaluation

The essence of technology testing is to fully focus on the benchmarking technology in the background of the system and to neglect errors in other parts and errors coming from other sources. Often it is testing using pre-prepared data in laboratory (ideal) conditions – meaning, there is no influence of the behaviour of any subject related to momentary feelings or system feedback. Technology testing is usually easily repeatable. The only limitation to this approach is acquiring a suitable database.

Scenario evaluation

The essence of scenario testing is to focus on the system as a whole. Testing focuses not only on the technology but also on the quality of other parts of the system. However, proper deployment and operating conditions also play an important role in the outcome of testing.

During the scenario testing, human subjects that may be used for the purpose of simulating the real-world situation can influence the system's behaviour unwillingly, by momentary feelings or inconsistency in using the system, or willingly by observing the system feedback and purposely changing their system usage.

All variable aspects related to subjects must be directed and recorded. Depending in the data used, this approach is partially repeatable.

A separate mention is dedicated to testing the presentation attack detection systems (abbreviated as PAD). Some biometric systems contain PADs as subsystems to detect attempts of presentation attacks, such as liveness detection systems. These can be tested separately from the rest of the system. The main requirement, however, is the ability to record their output directly.

3.3.2 Test design

The first step of any testing is design. The design establishes the general testing procedure, i.e. whether only the technology is tested or the entire product in a specific deployment scenario. Planning also includes gathering the information about tested system (logging, system feedback, ...) and the use-cases that dictates the potential testing scenario. Next, given the methodology, the data collection process is designed. Typically, this is either database acquisition for technology testing purposes or recruitment of test subjects for scenario testing. Finally, metrics for the measurement phase are proposed based on the set goal.

3.3.3 Test corpus

As previously indicated, test samples – data – are needed for testing purposes. For technology testing purposes, existing biometric sample databases may be sufficient, or custom databases may need to be created as part of the testing. The process of gathering the samples needs to be strictly directed and recorded. This typically takes place in laboratory conditions.

However, in case of scenario testing, it is also possible to use existing datasets that contain data from similar scenarios for this purpose, or artificially modified data to simulate a particular scenario. In the event of using real human subjects, it is recommended to gather a group of people similar to the target audience and record thoroughly the whole process of subjects using the system.

3.3.4 Measurement

During the measurement of results when testing a biometric verification system, the measured values need to be appropriately represented and evaluated. For this purpose, during the design phase of testing, the metrics on which the testing focuses are specified. Typical metrics recommended by the [17] standard are based on the technology/scenario approach.

Technology metrics

FMR (*False match rate*) and FNMR (*False non-match rate*) are typical metrics used for technology evaluation. They symbolize the rate of attempts that are falsely evaluated as genuine and those that are falsely evaluated as not genuine. These metrics only include the errors of matching algorithm. A specific threshold is required to be able to compute them. The relation between the metrics is:

$$FMR = 1 - FNMR$$

Scenario metrics

As for scenario-related metrics, the first metric is FTE (*Failure to enroll*). This metric shows the rate of failed attempts to enroll as new system user. Second commonly used metric is related to sensors – FTA (*Failure to acquire*). This metric is used for evaluation of rate of attempts to acquire data samples from user.

Combined with previously presented metrics FMR/FNMR we get a scenario-testing specific metrics – FAR (*False accept rate*) and FRR (*False reject rate*). These metrics include FMR/FNRM as well as FTE and FTA. The metrics symbolize the rate of falsely accepted users and falsely rejected users. FAR and FRR also require a specific threshold. A visualization of these metrics are in Figure 3.3.

When it comes to evaluating the presentation detection systems, standard ISO/IEC 30107-3:2017 [18] proposes a set of specific metrics to evaluate them. There are two metrics related to performance of the PADs – APCER (Attack presentation clasification error rate) and BPCER (Bona fide presentation error rate). They are evaluated based on the direct output of the PAD subsystems. The equations for computing are as follows:

$$APCER = 1 - (\frac{1}{N_{PAIS}}) \sum_{i=1}^{N_{PAIS}} Res_i$$

where:

- N_{PAIS} – number of presentation attacks for the PAI (Presentation attack instrument) types
- Res_i – the result from PADs (0 for not classifying as presentation attack or 1 otherwise)

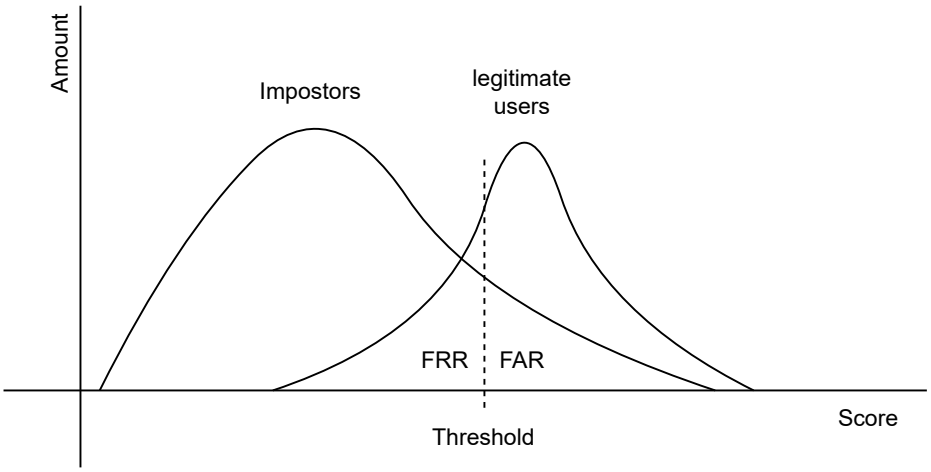


Figure 3.3: FAR vs. FRR metrics visualization

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}$$

where:

- N_{BF} – number of bona fide presentations
- Res_i – the result from PADs (0 for not classifying as presentation attack or 1 otherwise)

ROC/DET curves

As mentioned above, most metrics need a specific threshold to be able to tell, whether the attempt was accepted or rejected. But some biometric systems do not have such threshold set by the developing company, but rather by the clients. In this case, it is possible to plot the results into a curve. The curves show the overall performance of the system for a wide range of possible thresholds. ROC and DET curves are a different interpretation of the same thing, thus they are interchangeable.

As this work uses only the ROC curves, I will focus on describing them based on [3]. The ROC curve plots two metrics against each other – TPR (True positive rate) and FPR (False positive rate). Each point on the curve stands for one potential threshold – increasing the threshold decreases the FPR but also the TPR. The closer the curve to the center (45 degrees line) the worse performace.

3.3.5 Report

The conclusion of each biometric system testing is an output report. It contains the procedure according to which the test was carried out. Furthermore, the measured values in a suitable form. Finally, it contains a conclusion, i.e. an evaluation of the state of the system based on the measured data of each metric and any other relevant information.

3.4 Phonexia

The thesis focuses on a particular voice biometric verification system provided specifically for this purpose – Phonexia. This section contains basic information about the system, common use-cases and basic description of provided products. All information in this section is retrieved from official product website [2].

3.4.1 About Phonexia

Phonexia is a Czech software company focused on developing biometric system for voice verification, speech-to-text transcription and language, gender or even age recognition. In 2021, Frost & Sullivan acknowledged Phonexia in its report and Phonexia also won 2nd place in VoxCeleb Speaker Recognition Challenge 2021 in its respected category. Phonexia is world-renowned for quality of their product, even now used by the German Federal Criminal Police. Technical details about provided system will be discussed later in this work.

3.4.2 Common use-cases

Call centers

One of many use-cases, as suggested by Phonexia, are call centers. Benefits resulting from the use of such system could enhance advertisement targeting thus profit. Call center employee can see all the relevant information about the identified client in real time during their call, enhancing both client experience and company profits.

Remote identity verification

Another example of a proposed usage of Phonexia system is in the field of remote identity verification. With recent home-office working trends, a way to verify workers identity became much more desired by companies. Phonexia offers a system to verify or identify remote employees using only voice, as well as allow clients to access applications or private data only using their voice.

Besides verifying the workers, some industries are in contrary interested in their users. Banking and financial services are fondly using voice as another layer of client data protection. Phonexia also allows using voice for a fast authentication during a call without a need for password, including the fraud detection mechanisms.

Forensics

Phonexia product is actively used in different fields too. The German Federal Criminal Police uses Phonexia Voice Inspector for forensic analysis of the evidence to determine whether the audio recordings are forgeries or not. For this purpose, Voice Inspector offers automatic, language independent voice analysis and comparison. Besides that, a in-built wave editor as well for appropriate recording editing.

3.4.3 Provided products

Speech platform

Phonexia Speech Platform is a general software solution to cover Phonexia components and unite them into one product, customizable for clients needs. The platform itself can be divided into three components – Speech Engine (the core component that contains all the speech-related technology, such as voice biometric system or transcription system), Browser (graphical application to work with Speech Engine and visualize results) and utilities (RLS – Reporting and Licensing Server).

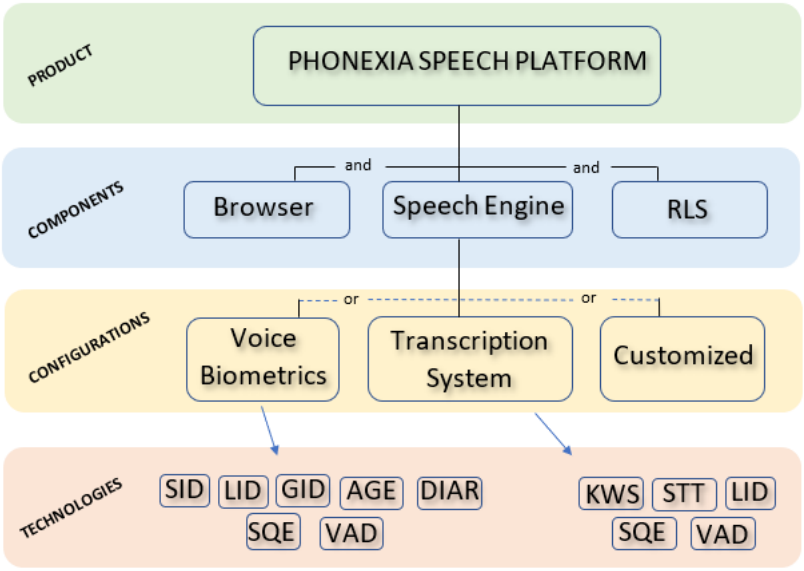


Figure 3.4: Phonexia Speech Platform architecture. Image retrieved from [2]

Voice Inspector

Next up is a Phonexia Voice Inspector. Voice Inspector is a specialized tool for forensic experts and police. The tool aims to be more precise than the general solutions with great support for voice analysis and recording editing. Besides automatic analysis, the tool also helps with creating reports for the court.

Voice Verify

Phonexia Voice Verify is a product specialized for use in call centers. The tool comes as API for integration into existing architecture. It provides technology for enhancing security alongside with tools to reduce call handling time.

Orbis

Phonexia Orbis is a specialized tool for law enforcement agencies. It is designed for fast investigation of audio files, visualization of analysis and report creation. In addition to the voice technologies, the tool includes a smart audio player with speaker highlighting, network maps for visualizing relations between people and assets and advanced user/case management system.

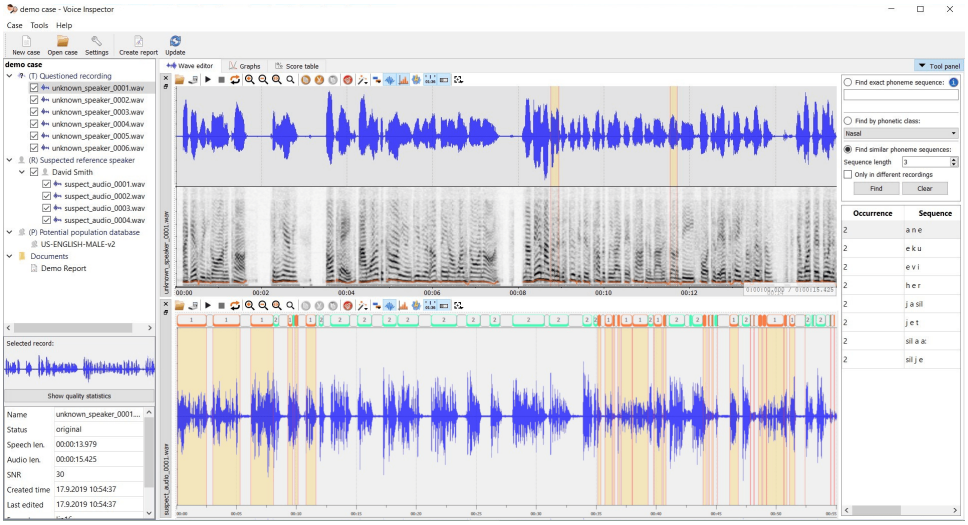


Figure 3.5: Phonexia Inspector interface. Image retrieved from [2]

Chapter 4

Testing method

Upcoming chapter covers the method for testing the robustness of provided Phonexia system against deepfake spoofing attack. The proposed method was put together with the idea to be close to the potential real attack, as is possible today, while still be as close as possible to the existing standard for testing the biometric systems – ISO/IEC 19795-1:2006 [16], ISO/IEC 19795-2:2007 [17] and ISO/IEC 30107-3:2017 [18], while incorporating the focus on deepfakes.

First section of this chapter describes the planning of the test – defining the attacker model to get the overview of the attacker point of view, setting the goal of the test, determining the use-case of system under test and summarizing everything into test scenario. Second section talks about selected dataset for testing. The last section talks about metrics for evaluating the measured results.

4.1 Planning the test

First phase of any system evaluation is about information gathering and setting the necessary objectives. During the planning phase, I constructed the attacker model, set the goal of testing, determined the use-case of the target system to test and combining those aspects into a test scenario.

4.1.1 Defining attacker model

Before any test method drafting, it is important to realize and note, what are the actual threats to your system. It is necessary to be aware of threats before constructing a real test scenario. A way of realizing this is creating a threat model – specifically a attacker model. First part will be about creating a attacker-centric threat model according to OWASP[1] guidelines [4]. This model summarises what is the attacker's motive, means and opportunity to make an attack on our system. Second part rates the attack according the metrics proposed by Tekampe et al. [36].

Motive

There are many possibilities to consider when it comes to motive of potential attacker. Generally, the nature of attacker motive could be seen as a desire for acknowledgment (mental desire) or desire for valuables (material desire).

[1] https://owasp.org/

Transferred to real life examples:

- Desire for acknowledgment – by successfully breaking through the system, the attacker will gain fame in the hacker community
- Desire for valuables – by successfully breaking through the system, the attacker will either gain access to valuables of the victim or create some damages

In terms of the provided system this thesis works with – Phonexia, the real life examples depend on the use-case of the system (3.4.2. In the case of using the system as a identification tool, attacker could trick the system to identify him as a different client and thus cause damages to the target client reputation or to the company itself by bypassing the identification step and make them unable to track the client.

As for the case of using the Phonexia system as verification tool, the attacker could trick the system to verify him as a client or employee and thereby grating access to, for example, client's banking account or the company infrastructure (including confidential documents) the employee works at.

Means

One of the scary sides of deepfake technology is the accessibility and learning curve for production tools. All the attacker needs is a reasonably fast computer and a deepfake generator. There is a number of open-source voice deepfake tools with wide community and documentation. Table 4.1 shows examples of open-source tools available.

Tool name	Source
MozillaTTS	https://github.com/mozilla/TTS
YourTTS	https://github.com/edresson/yourtts
CoquiTTS	https://github.com/coqui-ai/TTS
Real-Time-Voice-Cloning	https://github.com/CorentinJ/Real-Time-Voice-Cloning

Table 4.1: Voice deepfake tools

Once the attacker obtains a tool and learns how to create samples, the only thing that remains is sending the prepared recordings to the biometric system via any phone call software.

Opportunity

In real life scenarios, the Phonexia system is usually deployed in a environment, where restricting the access to the system input is not really an option. The products that integrates Phonexia are build to receive phone calls and redirecting the incoming voice to the Phonexia Speech Engine for analysis. That means that any attacker can get access right to the system with just a phone.

The main barrier when it comes to opportunity to make an attack highly differs based on the use-case of the system. Since for identification bypass, the only thing a attacker needs is any kind of deepfake. On the other hand, if attacker wants to impersonate another client, he needs a source of the victims biometric – voice recording.

It is not a big problem to get voice recordings of a large number of people who actively and carelessly use social networks where they share videos including audio recordings. The only challenge is to acquire speech long enough for the tools to be able to produce forgeries

of sufficient quality. That means, that the opportunity for an attack is created by the clients themselves.

Rating an attack

Second part is dedicated to the description of the attack rating proposed by Tekampe et al. [36] and evaluating the possible attack on the Phonexia system.

The most important part of the attack rating is the rating table. The table contains sub-factors and their ratings to identify an attack on the system and its exploits. As indicated in the table B.1, there is a difference between identification and exploitation. The main difference, as described in [36] is as follows. Identification refers to the effort required to discover an attack and demonstrate it in both laboratory and real-world deployments. Exploitation, on the other hand, is the effort required to successfully execute an attack on a system according to the analysis and procedure from the identification phase.

Next up is a brief description of sub-factors presented in table B.1 in appendix B:

- Elapsed time – The elapsed time factor expresses the time required to perform an action. In the identification phase, this factor indicates how time-consuming it is to identify the attack – that is, to discover, demonstrate and write the necessary texts to reproduce the results. For the case of a presentation attack (spoofing), this includes, for example, the time spent searching for the so-called *Golden Fake* [36]. For the exploitation phase, it means the time needed to reproduce the result of the identification part in a real environment.

- Expertise – Expertise is a factor that expresses the attacker's level of ability to perform a successful attack and general knowledge. The rating proposed in [36]:

  Layman – no special experience, knowledge or skills are required, a general education is sufficient

  Proficient – knowledge of the field (biometrics), knowledge of existing attacks and possible basic adaptation of procedures for the specific case is required

  Expert – specific attack preparations are required and possibly also the attack know-how itself

  Multiple experts – it takes a group of experts from different sectors to attack

- Knowledge of system – The factor expresses the level of system knowledge required for a successful attack. For example, it can be the product architecture, communication protocols or data format. The proposed rating according to Tekampe et al. [36] is as follows:

  Public – information about product is publicly available

  Restricted – information about product is available only to developers and partners under NDA

  Confidential – information about product is not shared outside of the product company

  Critical – information about product is only shared among specific people

- Access to the system / Windows of opportunity – A factor expressing the difficulty of accessing the system. In the context of identification, this is the difficulty of find-

ing/purchasing the system for testing and any other equipment needed. For exploitation, it is the difficulty of accessing the deployed system and bypassing any additional security. The evaluation according to Tekampe et al. [36] is as follows:

Easy – product is easily accessible and attacker has unlimited number of attempts to attack

Medium – the product is limited, not accessible to individuals, attacker has limited number of attempts

Difficult – product is available only for identified users, exploitation requires very specific settings and often the cooperation of people in the target company

- Equipment – This factor expresses the level of equipment needed to attack. According to Tekampe et al. [36], it includes biometric databases. The proposed rating is as follows:

  Standard – common equipment, no specialized tools needed, easy to obtain or make

  Specialized – expensive tools, hard to obtain, equipment for specialized tasks

  Bespoke – equipment with restricted access, very expensive tools

- Access to biometric characteristics – The biometric access factor expresses the difficulty in obtaining biometrics to attack the product. In the case of a presentation attack, the original is assumed to exist and its acquisition and production of a forgery are evaluated. The evaluation according to Tekampe et al. [36] is show in Table B.2.

- Resistance evaluation – The overall system resilience is calculated by summing the corresponding values from the B.1 table. The Table B.3 then corresponds to the level.

Next comes the evaluation. Upcoming table shows the values of each factor and the final score.

Factor	Identification	Exploitation
Elapsed time	2	2
Expertise	2	0
Knowledge of system	0	0
Access to the system	2	0
Equipment	0	0
Access to biometric characteristics	0	0
Evaluation	6	2

Table 4.2: The results of attack rating

The final score is 8, which according to table B.3 means, that the system provides no resistance to presentation attack using deepfake, assuming that the attack can be successful (a tool exists to create deepfakes of sufficient quality).

Assumptions made during the evaluation:

- Elapsed time – the tools and quality of deepfake production is not evident, the attacker needs to try multiple times, which might require up to one month of tools research and creating recordings of sufficient quality. In the exploitation phase, the attacker possibly needs more than one day to recreate the attack but no more than a week

- Expertise – for the identification, a expertise is required to be able to operate and modify tools. No special expertise is needed for recreating the attack
- Knowledge of system – all documents needed are publicly available online
- Access to the system – access is through phone calls, thus almost unlimited
- Equipment – only basic equipment needed for creating deepfakes and accessing the system
- Access to biometric characteristics – according to table B.2.

4.1.2 Setting the goal

First step of any systematic testing is declaring, what the goal of this particular testing is. Usually the tester decides, what aspects of the tested system are the main focus.

The goal of a general testing of system robustness against an attack is to acquire statistical data on system resistance/vulnerability to a given attack. In the case of this thesis, the provided system is tested against a presentation attack, described in 3.2.2.

As for this test, I will try to measure the robustness of provided biometric system, Phonexia Speech Platform 3.4, which will be deployed as solitary system (without the usual integration into existing products), against different groups of deepfakes with the main goal of determining the resistance to different methods of creating forgeries.

4.1.3 Determining system use-case

Now, I as a tester, have set the goal and outline of the testing method. Next up is deciding, which of the possible use-cases of the provided system will be used.

There are plenty of available use-cases, some listed in the section 3.4.2. From a tester's perspective, the use-cases are divided into the groups according to the main function of a system (mode of comparison):

- verification – comparing the input samples and stored templates (*voiceprints* in the case of Phonexia) 1:1 – the user declares who he/she is and then presents the proof in the form of biometric
- identification – comparing the input samples and stored templates (*voiceprints* in the case of Phonexia) 1:M – the user does not declare who he/she is and presents the biometric, the system then tries to identify the user by matching the input sample to the existing templates in its database

This test will focus on one of those described groups of use-cases – verification. During the experiments, the Phonexia system will be treated as a solution for companies requiring clients or employees to verify themselves remotely using voice as a password.

4.1.4 Test scenario

According to standard IEC/ISO 19795-2:2007[17], there are two general approaches to testing a biometrics system – technology evaluation and scenario evaluation. As described in 3.3.1, the methodology of technology evaluation focuses on testing the internal forgery

detection methods, where as the scenario evaluation focuses on testing the system as whole product. Testing all thinkable aspects is out of scope of one work.

The thesis will look at the biometric system, and whole scenario, from the view of a potential attacker. Although the complete Phonexia Speech Platform product, as show in 3.4, is deployed specifically for the purpose of this work and available during the process of testing, the system will be considered a black box [10]. The provided system will be used as a voice-as-a-password tool, simulating the remote verification for clients.

Since the Phonexia Speech Platform is designed to be integrated into a existing system, the product itself isn't structured as a common biometric system. As the main idea behind Phonexia system is remote identification, verification or analysis of audio recordings, the system does not come with any sensors (microphone) – the main input is either audio file or audio stream. This aspect reflects into the data selection discussed later in 4.2.

As suggested in [10], the general evaluation protocol for black box testing of a biometric system can be divided into two phases:

- sample collection – in the context of this work, this is the phase of obtaining a suitable dataset
- cross-comparison – phase of presenting the samples to the system and collecting the result

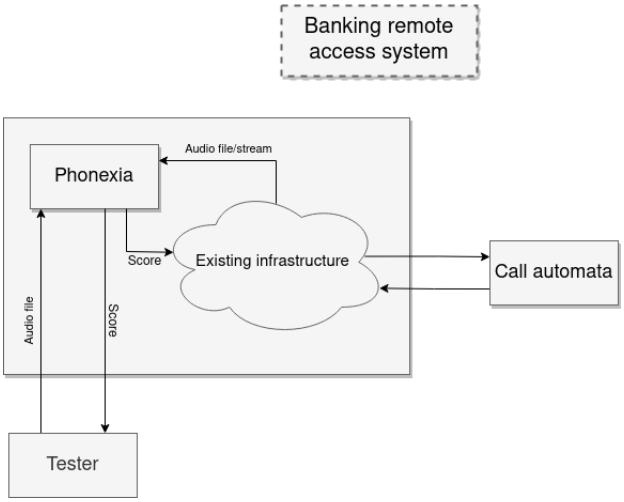


Figure 4.1: Logical architecture scheme

4.2 Dataset

Coming next is the section about sample collection phase. First discussed will be the data properties – the frequently mentioned ones as well as the ones potentially important. Followed by a description of collected data from publicly accessible datasets, listed in 2.4, and reasoning behind the choice.

4.2.1 Data properties

IEC/ISO 19795 [17] describes the process of data/samples collection vaguely. For the case of real-world scenario system evaluation, the data collection process is summarized as hiring

as many test subjects (people) as possible. This approach gives developers the idea whether there is a problem with their product or not.

What this work aims to achieve is to not only test the provided system, but to also try to specify, what may have caused such behavior in terms of deepfake variety. No standard or studies have been found to consider test subject characteristics or characteristics of spoofs creation. Such data properties could have indirect impact on the biometric system performance. Data properties to consider could be as follows:

- language – for the sake of testing the impact of language on the biometric system's detection
- sex – evaluation of the impact of sex of the subject (possible impact on the deepfake generation as well)
- recording environment – usually studio or any real-world scenario (original samples collected from social media)
- audio quality – potentially high impact on the results, this property refers to audio noise caused by either capture device quality or adding artificial noise for the sake of capture device simulation
- deepfake quality – currently a widely discussed topic in the deepfake security community, the deepfake quality still isn't measurable at the time, although progress is pushing through in the more popular area of face deepfakes [1]

My work will focus on the deepfake aspects. Deepfake quality, as suggested in the overview, is currently a scientific topic with no exact results. As for this work, I will not rate the deepfake quality but rather differentiate between the different deepfake methods of creation and consider them to generate constant level of quality samples.

4.2.2 Selected data

As specified in the assignment, the deepfake datasets available online must be used to test the provided product. Some of the currently available deepfake datasets are listed and are briefly described in the section 2.4. Here the thesis focuses on the selection of a dataset, specific data and their properties according to the suggested scenario.

In addition to the mentioned properties of the data, the number of samples used to perform the test must also be considered. Neither standards nor studies on this topic give a specific number needed. The general recommendations as given in [17] encourage as many test subjects as possible. The number of samples does not matter to the testing itself, but rather to the interpretation of measured results and their relevance.

Given the available information on datasets and their availability, this work will use the aforementioned ASVspoof2019 dataset for its variability in data quality (i.e., the division into the two categories of LA and PA as outlined in 2.4) and natural fitness for demonstrative testing.

The ASVspoof2019 dataset consists of two sections – PA and LA. As already described in 2.4, those stand for Physical Access and Logical Access. In the Figure 4.1 we can see the logical architecture of tested system. Considering that the only available part to test is the Phonexia subsystem, only the LA part of dataset will be used for testing. PA part would be usable too, but since the goal is to test the difference among the deepfakes methods, not the audio quality, and looking at the fact that the infrastructures preceding Phonexia

subsystem could often contain noise filtering as well, by eliminating the additional noise from data I'll ensure more consistent results.

4.3 Evaluation

The following section addresses the measurement phase. The primary focus of this section is on describing the commonly used metrics according to the standard for evaluating the robustness of biometric systems, as well as describing the metric used to evaluate deepfake detection methods as a way to evaluate robustness according to each method of creation as suggested in the 4.1.2 section.

4.3.1 Metrics

Common metrics for measuring the performance of a biometric verification system can be sorted into essential and less important/irrelevant from a perspective of this thesis. The first category discussed is irrelevant metrics. There are two metrics in total – FTE (Failure to enroll) and FTA (Failure to acquire).

FTE is a metric indicating the ability of the system to enroll a new entity without error. This characteristic is irrelevant from the perspective of an attacker on a system seen only as a black box. In the scenario tested, i.e., applying a presentation attack with the goal of deceiving the system into verifying attacker as a target client, the attacker does not encounter the registration phase, as he impersonates already successfully registered users.

The second irrelevant metric is FTA, which indicates the ability of the system to successfully sample a subject. This metric is unmeasurable in this scenario due to the nature of the architecture used (the registration samples would most likely be prepared for Phonexia by the existing infrastructure of a company) and the principle of test execution, as individual samples are not taken as part of the test, but only mined from previously taken datasets. Therefore, in principle, it is not possible to consider faults in the sensors or any other component of the sample collection module.

The second category is important routine metrics. These are the FAR (*False accept rate*) and the associated FRR (*False reject rate*). Their general relationship can be expressed as $FAR = 1 - FRR$. The metrics express the proportion in which the system incorrectly accepts fraudulent acceptance attempts (FAR) or the proportion in which the system incorrectly rejects legitimate acceptance attempts (FRR).

The FAR and FRR metrics are often confused with their counterparts FMR (*False match rate*) and FNMR (*False non-match rate*). However, the slight difference between them is relevant to this paper. The main difference is in the relation of these metrics to the previous ones. While FMR shows purely the decision outcome of the system, in FAR the FTA metric is also included. Since, as described, FTA is not relevant for this work, only the FMR and FNMR metrics will be considered subsequently. As part of the evaluation, these metrics will be displayed using ROC/DET curves. It is worth mentioning, that the FMR/FNMR metrics are commonly used in the technology evaluation, rather then scenario evaluation. But again, from the point of view of the architecture of a tested system, the data shown later in this work are based on these metrics.

The last mentioned metrics in 3.3.4 are APCER (Attack presentation clasification error rate) and BPCER (Bona fide presentation error rate). These metrics are presented in standard ISO/IEC 30107-3:2017 [18] for representing the ratio of presentation attack detection.

These metrics are unusable due to the black-box testing. These metrics required direct output from PAD systems (Presentation attack detection).

As stated earlier, in addition to the standard metrics for measuring the robustness of biometric verification systems, this thesis will also present other metric. This will be metric used in describing the performance of deepfake detection methods – the AUC (*Area uder curve*) [21], sometimes also called AUROC (*Area under ROC*) [42]. As the name suggests, the metric gives the percentage area under the curve. It is a simplified evaluation of ROC/DET curve plots. AUC is mostly used as a simplified representation of ROC curves that allows rating the system performance base on single number.

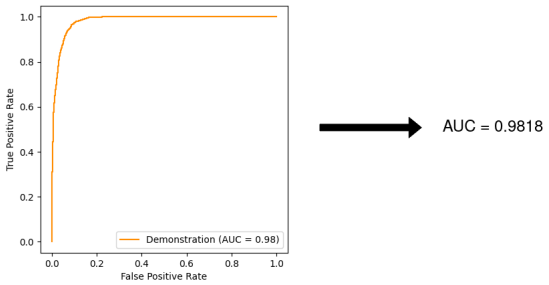


Figure 4.2: ROC vs. AUC

In addition to these general metrics, other metric is introduced in this thesis to discover the properties of the input data that could significantly affect the decision making results of the verification system.

The metric is *AUC vs deepfake type*. This metric is inspired by custom metrics proposed for evaluating the detection method in [22]. *AUC vs deepfake type* is a metric expressing the relationship between AUC and the audio deepfake type. As suggested in 4.2, the importance of this metric lies in differentiating between methods of creating the deepfakes and their impact on deepfake strength against the Phonexia system.

The main benefit of this metric is splitting the ROC curve into smaller, more specific ones. In case of only one ROC curve for all methods of creating deepfakes, some potentially very strong methods could remain hidden while posing a threat to the system.

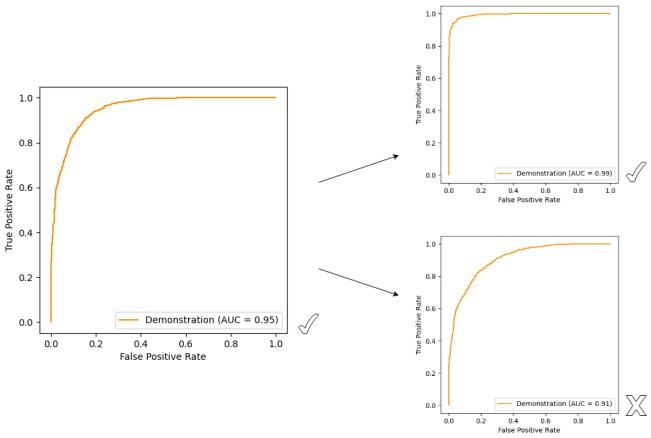


Figure 4.3: Example of *AUC vs deepfake type* metric where the company with 0.95 AUC limit would miss potentially dangerous method

Chapter 5

Conducting the tests

This chapter describes the experiment based on the previous design.The first section briefly describes the environment of the experiment. Next, the work describes the features and functions of the Phonexia system, including the procedure for using the described functions during testing. Then, the details of the experiments are described – how the experiments fulfill the designed scenario, how the data from the selected dataset was used, and what the measured values look like. The second to last section is the evaluation of the measured data. The last section is dedicated to a brief discussion of possible improvements to the system based on the results.

5.1 Environment

The first section is a description of the environment and deployment of the tested system. The deployment of the Phonexia Speech Platform system is typically in the hands of clients in a real-world environment. They receive a license and integrate the delivered system into their existing infrastructures. The Phonexia system itself therefore does not have a precise deployment and usage procedure.

The delivered system was therefore deployed as a stand-alone product in a virtual machine without any surrounding infrastructure. The system can be interacted with either via a graphical browser (when connected via remote desktop – not usable for automated testing) or via a REST API.

As already mentioned, the system was deployed as a stand-alone product, i.e. without a surrounding application that would process the output scores and that would also supply input samples (since Phonexia is a software product and therefore does not come with sensors). The logical scheme of the deployed system is illustrated in Figure 5.1.

5.2 Phonexia Speech Engine

The following section describes the relevant properties of the system. In particular, the focus is on the input data requirements for the proper functioning of the system. Next, the system's REST API, the system functions used and the procedure for using them during testing are described.

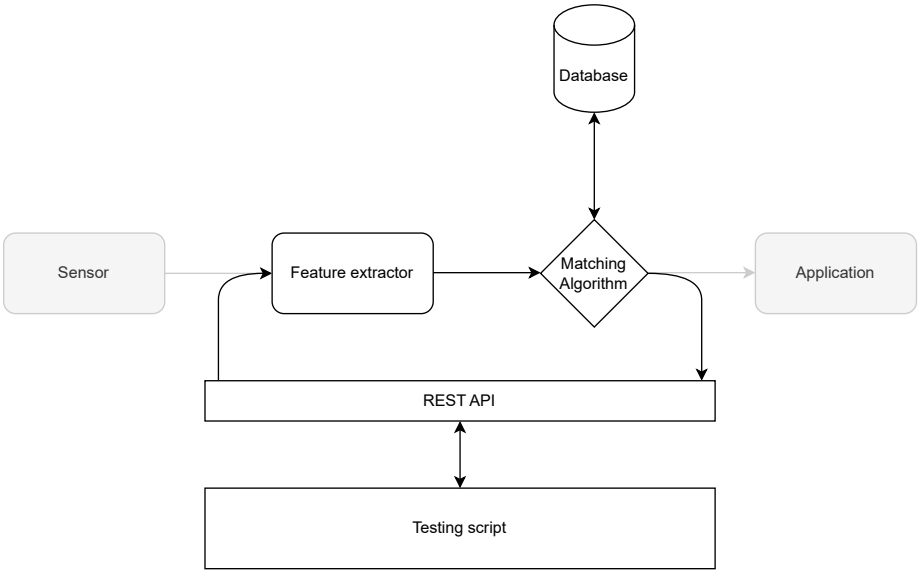


Figure 5.1: Logical scheme of the test environment

5.2.1 System properties

The first feature of the system is a logging of the transactions. According to standards, it is required keep information about the transactions made. If the system does not store it in its own database, it needs to be recorded manually. Phonexia does not have this capability, so transactions need to be logged by testing script. Sufficient information will be kept during testing to reproduce the tests.

Another property is, again according to the standards, the logging of results. As with transactions, results need to be stored for each attempt. Phonexia does store the results of individual queries, but only in a temporary cache that is set up for asynchronous requests. Thus, during testing, results will be recorded continuously alongside transactions.

As for the results themselves, Phonexia Speech Engine returns a score as a result of the comparison. This score is the log-likelihood ratio. Which means that it is theoretically possible to get values from $-\infty$ to $+\infty$. In real use, however, this is typically a range of approximately -30 to $+30$. The system is calibrated to a base value of 0, which should mark an imaginary threshold, but in practice it is often necessary to shift this value according to the data being used.

Another important property of biometric systems is the updating of user templates. Systems that support this feature update the template after each sample is accepted according to the input data received. However, Phonexia does not support this functionality. Saved templates are unchanged regardless of the test result.

The last, and probably most important part, is the system's input data requirements. Typically, biometric systems will check important properties of the data when taking a sample and based on these, accept the sample as valid or reject it with a Failure to Acquire error. Phonexia, although it has specified requirements, does not check these properties. It leaves the checking process to the assumed surrounding application. It is therefore important to ensure these properties during testing.

The first such property is audio quality. In real traffic, it is common to receive data with excessive noise, which prevents reliable user authentication. And although Phonexia

provides an audio checking tool for this purpose, its use in the proposed scenario and with regard to the data used is not necessary, as part of the dataset used contains recordings without significant quality degradation. However, if Phonexia receives a sample with poor audio quality, it will give it a low score indicating non-acceptance.

Besides audio quality, Phonexia Speech Engine also requires a certain minimal length of speech in the proposed samples. Based on information provided directly by Phonexia developer team, a minimum of speech required for enrollment of new user is 20 seconds (for the use-case of verification system). As for the verification itself – 3-5 seconds of speech is enough.

5.2.2 REST API

Functions provided via Phonexia Speech Engine REST API are divided into the areas of system technologies. First are are basic operations – authentication and file manipulation. Second are the function specific for Speaker Identification – a technology to identify (verify) a user in provided recording. At the end, there is a description of how the functions are used during the test.

Basic operations

Login operation. Used for user authentication to Phonexia Speech Platform. After successful login using username and password, session ID is returned. The session ID is required by other requests (header parameter `X-SessionID`).

```
Login (POST):
/login

Headers:
Authorization: Basic *username:password encoded in base64*
```

Audio-file upload operation. Used for uploading an audio recording into the directory specified by `FILE_PATH` of logged user. After successful upload the audio recording information is returned.

```
Upload audio-file (POST):
/audiofile?path=/

Headers:
X-SessionID: Session ID returned by /login
Body:
Audio-file (Only for upload)
```

Audio-file delete operation. Used for deleting an audio recording from the directory specified by `FILE_PATH` of logged user.

```
Delete audio-file (DELETE):
/audiofile?path=/*FILE_PATH*

Headers:
X-SessionID: Session ID returned by /login
```

Speaker Identification

Create new speaker model operation. Used for creating new empty speaker model. Speaker model name is specified by `SM_NAME`. Speaker model stands for user profile that is used for identification/verification.

```
Create speaker model (POST):
/technologies/speakerid4/speakermodels/*SM_NAME*

Headers:
X-SessionID: Session ID returned by /login
```

Delete speaker model operation. Used for deleting speaker model including all audio-files uploaded into this specific speaker model. Speaker model name is specified by `SM_NAME`.

```
Delete speaker model (DELETE):
/technologies/speakerid4/speakermodels/*SM_NAME*

Headers:
X-SessionID: Session ID returned by /login
```

Audio-file upload operation. Used for uploading an audio recording into the directory specified by `FILE_PATH` of speaker model specified by `SM_NAME`. After successful upload the audio recording information is returned.

```
Upload audio-file into speaker model (POST):
/technologies/speakerid4/speakermodels/*SM_NAME*/audiofile?path=/

Headers:
X-SessionID: Session ID returned by /login
Body:
Audio-file (Only for upload)
```

Audio-file delete operation. Used for deleting an audio recording from the directory specified by `FILE_PATH` of speaker model specified by `SM_NAME`.

```
Delete audio-file from speaker model (DELETE):
/technologies/speakerid4/speakermodels/*SM_NAME*/audiofile?path=/

Headers:
X-SessionID: Session ID returned by /login
```

Prepare speaker model operation. Used for creating voice-print of speaker model specified by `SM_NAME` (biometric template) thus preparing said speaker model to be used for identification by using pre-trained model specified by `MODEL`.

```
Prepare speaker model (PUT):
/technologies/speakerid4/speakermodels/*SM_NAME*/prepare?model=*MODEL*

Headers:
X-SessionID: Session ID returned by /login
```

Identify speaker operation. Used for identifying/verifying speaker specified by his/her speaker model name `SM_NAME` in uploaded audio recording specified by `FILE_PATH` using pre-trained model specified by `MODEL`. After successful comparison of the voice-prints a score of comparison is returned.

```
Identify speaker (GET):
/technologies/speakerid4?path=/*FILE_PATH*&
    speaker_model=*SM_NAME*&model=*MODEL*

Headers:
X-SessionID: Session ID returned by /login
```

Complete procedure

As for a complete procedure – I followed the suggested examples of using the system by Phonexia documentation [5]. For the purpose of testing I used provided pre-trained model `XL4`. The complete procedure is as follows:

1. Login – Using `/login` to authenticate
2. Create speaker model – Creating new empty speaker model
3. Upload recordings into speaker model – Uploading bona fide samples into the created model
4. Prepare speaker model – Preparing model for comparison
5. Upload recording to analyse – Uploading spoofed sample
6. Speaker identification – Comparing the voice-prints of speaker model and spoofed recording
7. Cleanup – Removing audio recording and speaker model

Depending on the number of tested samples, the steps 5-6(7) were repeated. For multiple sample tests the procedure was to upload new spoofed sample, compared it to the existing speaker model, remove the spoof and repeat. After testing all spoofs for specific speaker, the speaker model was remove as well.

5.3 Experiments

Upcoming section discusses three areas of conducting the experiment. First discussed area is scenario fulfilment – a brief description of how the system was used to fulfil the proposed scenario. Second is the part about how the data was used to satisfy system requirements. And at last there is a part about the results of testing and logged information format.

5.3.1 Scenario fulfilment

reference k bodům v scenario v kapitole 4 nadhodit dataset - LA pro logic access jako je popsáno v ASVspoof19 paperu použití fcí systému tak, aby šlo o verifikaci

Based on the proposed scenario in 4.1.4, the system was as a verification tool. The complete procedure of testing is:

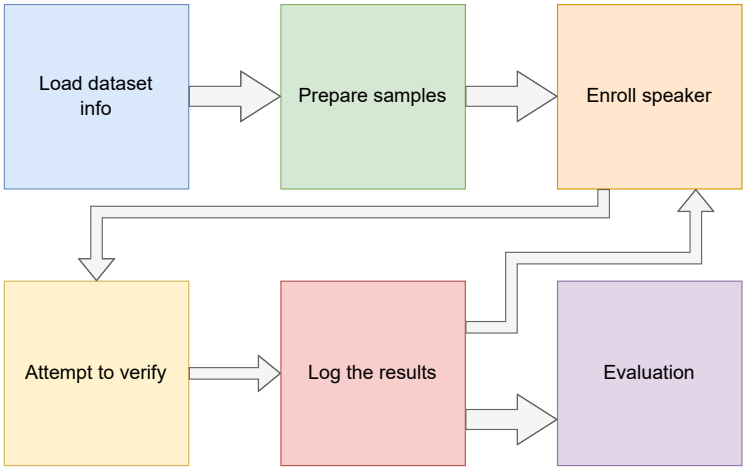


Figure 5.2: The test process workflow

1. Load the dataset information and extract the useful data
2. For each speaker in dataset, do following:

   Pre-process samples

   Get bona fide samples and enroll new speaker into the system

   Try to verify as said speaker using bona fide samples that belongs to the speaker (exclusive to the samples used for enrollment), using bona fide samples that does not belong to the speaker and spoofs of the speaker

   Clean up – remove files from the target system and remove speaker model

This procedure is simulating the existence of register user and the attacker's attempts to verify as the specific user.

5.3.2 Data usage

Next part is about using the dataset. As mentioned in 5.2.1, the Phonexia Speech Engine has strict requirements for input data for the use-case of voice verification system. The dataset used in this work (ASVspoof2019) however does not fulfil the required data properties.

The average length of a bona fide recording in used dataset is $2.59s$ and $3.79s$ for spoofed ones. To ensure the required speech length, considering the speech is approximately $\frac{2}{3}$ of the audio length, I was compelled to concatenate the samples into a longer recordings. All used samples were concatenated into one of five utterances.

For enrollment, the required length is set to $20s$ of speech. In total, I used all available bona fide recordings for each speaker, divided them into two groups of the same size – one for enrollment and one for testing.

For testing, the required length is set to $3 - 5s$ minimum. Again, to ensure the system functionality I concatenated the spoofs into one of five utterances with the regard to the method of creation. On top of spoofs and bona fide samples that belongs to the speaker (exclusive to the samples used for enrollment) I also added bona fide samples that does not belong to the speaker to find out the real scores for true reject.

Complete list of systems used for testing is shown in Table 5.1.

System ID	Description
A01	TTS system with WaveNet waveform generator
A02	TTS similar to A01 with WORLD vocoder instead of WaveNet
A03	TTS similar to A02. Can be built from scratch by TTS toolkit Merlin[1]
A07	TTS similar to A03 with GAN-based post-filter to mask artifacts
A08	TTS similar to A01 with faster waveform generator
A09	TTS system made for real-time mobile device synthesis
A10	TTS system based on Tacotron 2 reportedly with high naturalness
A11	Same as A10 except for Griffin-Lim algorithm waveform generator
A12	TTS based on AR WaveNet
A13	Combined neural VC with TTS
A15	Combined VC with TTS with WaveNet vocoders
A17	neural VC system

Table 5.1: Methods of creating the deepfakes in used dataset and brief description retrieved from [40]

5.3.3 Results

The results of the attempts were logged throughout the testing process. The recorded information is sorted by individual speakers. For each speaker part of log, the information is divided into two sections – information about file concatenation and section about the attempts.

Each line for the concatenation section is in following format:

```
==>TARGET_FILE_PATH: [LIST_OF_SOURCE_FILES]

TARGET_FILE_PATH - Path to the concatenated file
LIST_OF_SOURCE_FILES - comma separated list of concatenated files
```

Each line for the attempts section is in following format:

```
HASH: SPEAKERID FILENAME FILEPATH CHANNEL [SCORE]

HASH - unique hash
SPEAKERID - dataset specific ID for speakers (also used as speaker
            model name)
FILENAME - SYSTEMID_FILENUM_LENGTH_TYPE
    SYSTEMID - deepfake method ID (dataset specific)
    FILENUM - file sequence number in the context of testing script
    LENGTH - the total audio length
    TYPE - bonafide/spoof
FILEPATH - path to the file
CHANNEL - channel of audio used for comparison (in this case always 0)
SCORE - final comparison score
```

5.4 Evaluation

The penultimate section presents the measured results in the form of proposed metrics. It demonstrates the difference between using standard metrics versus custom metrics and what benefits it brings.

5.4.1 Metrics

The first step was to evaluate the robustness of the system to deepfakes using common metrics. This evaluation is shown by the ROC curve in Figure 5.3. This is a representation of the TPR versus FPR ratio as described in 3.3.4. As can be seen in the curve, the system appears to be very robust to the deepfakes dataset used.

The required level of robustness can normally vary by company and its target clientele and the intended use-cases of the system. For example, companies with products designed for employee verification that have access to highly sensitive data have high requirements for the AUC of their system. Even for these, however, a measured AUC = 0.99 may seem fully sufficient.

However, the next sequence of graphs 5.4.1 reveals the importance of focusing on individual deepfake generation methods. This is because in practice, an attacker will typically not try all methods for generating deepfakes, but will only select the best ones that have the highest chance of success. In this case, the system with identifier A10 could likely be the one in question. This, as can be seen in the curve and in the AUC table 5.2, has much more alarming results. For companies developing authentication systems for use in critical sectors, an AUC of 0.91 may already indicate the need for improved detection.

System ID	AUC	Evaluation
A01	1.0	OK
A02	1.0	OK
A03	1.0	OK
A07	0.99	OK
A08	1.0	OK
A09	1.0	OK
A10	0.91	?
A11	0.99	OK
A12	0.99	OK
A13	1.0	OK
A15	1.0	OK
A17	1.0	OK

Table 5.2: AUC vs. deepfake generation system ID

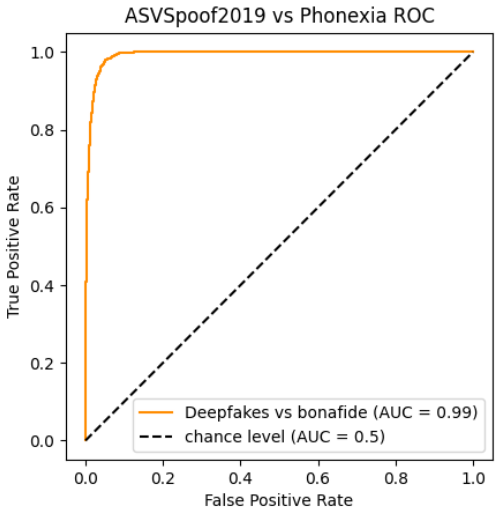
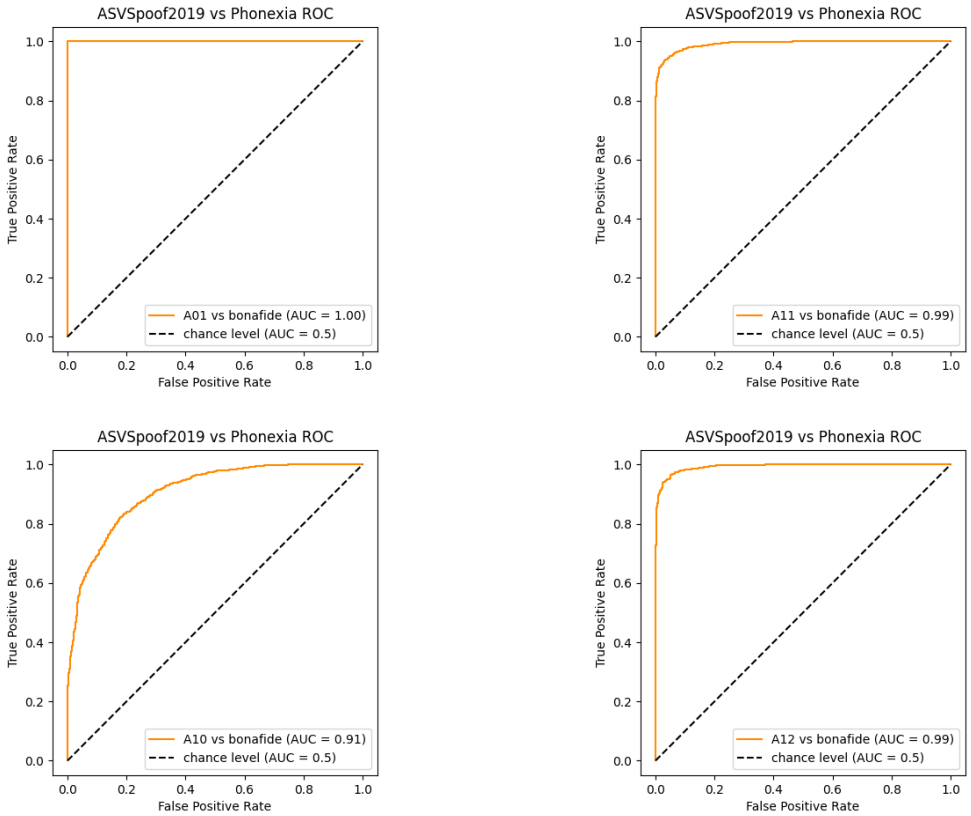
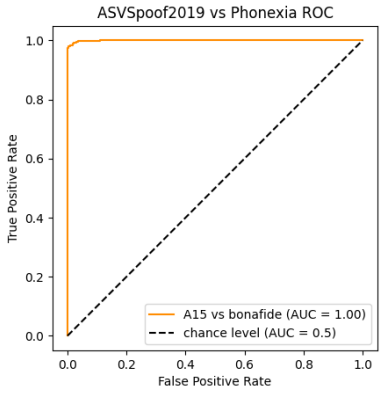


Figure 5.3: ROC curve of the system's performnace

Upcoming are the ROC curves for systems listed in 5.2. Same curve represents the A01, A02, A03, A07, A08, A09, A13, A17. At last there is a distribution graph 5.4 to display the overlap for systems A – bona fide attempts to the corresponding speaker, R – bona fide attempts to the non-corresponding speaker and A10 – the system with best performance in terms of breaching the system.





Graphs generated using code inspired by scikit-learn documentation examples[2]

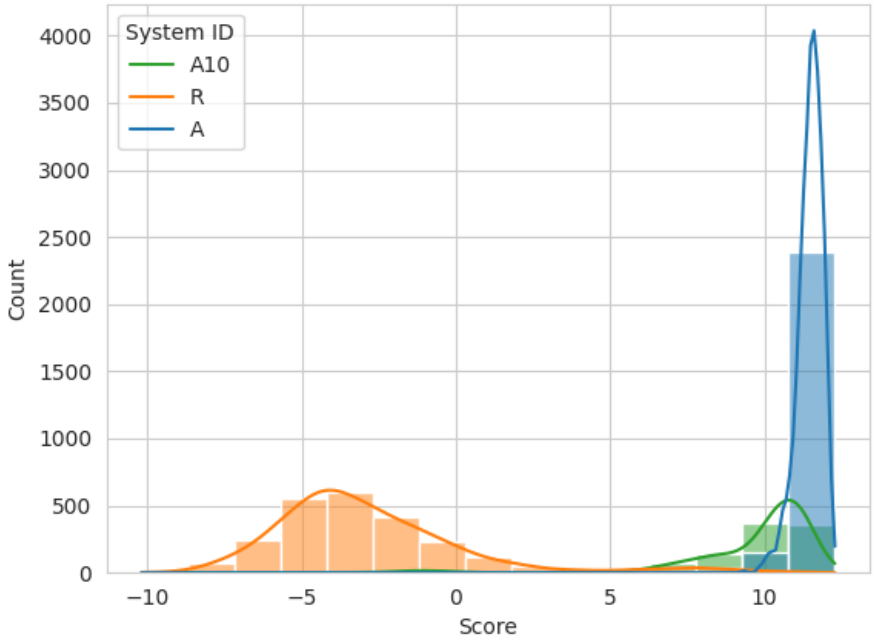


Figure 5.4: Results distribution graph

5.5 Improvements discussion

The last part is a short discussion on possible improvements to the system. The tested system shows generally high signs of robustness to the dataset used. However, it should be highlighted that for demonstration purposes a now older dataset was used.

Despite the relative outdatedness of the forgery generation methods, a system has been discovered that shows a higher success rate in breaching system protection. It can be

[2]https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

assumed that there will be more and more methods that are unsafe for the system as new methods continue to emerge.

For this reason, it is advisable to plan a scalable solution for detecting ever new types of deepfakes. The constant re-training of models can be a time-consuming and costly operation.

A possible solution could be a system of smaller detectors trained against a specific group of counterfeit creation methods. With the arrival of new deepfake creation methods, this would only require a smaller model to be retrained or created.

Chapter 6

Methodology

Upcoming chapter discusses the summary of previous steps into a general methodology. Goal of this step is to propose a repeatable procedure to follow during testing a biometric system against spoofing attack using deepfakes. As seen in Figure 6.1, the methodology has five main parts. Each of these parts will have it's own section.

Starting with planning phase – this section talks about planning the testing and what to consider during this step. Main areas discussed are identifying the system properties, defining the attacker model and system use-cases, determining the goal of testing and the summary into testing scenario.

Second section talks about dataset – the pros and cons of using the online available datasets or building your own. This section also talks about the properties of data and what to be cautious about.

Third section is about executing conducting the test – talking mainly about the environment and proper test behaviour.

Fourth and fifth sections are about the evaluation and interpretation. These sections discuss the metrics, data evaluation and the relevance of results.

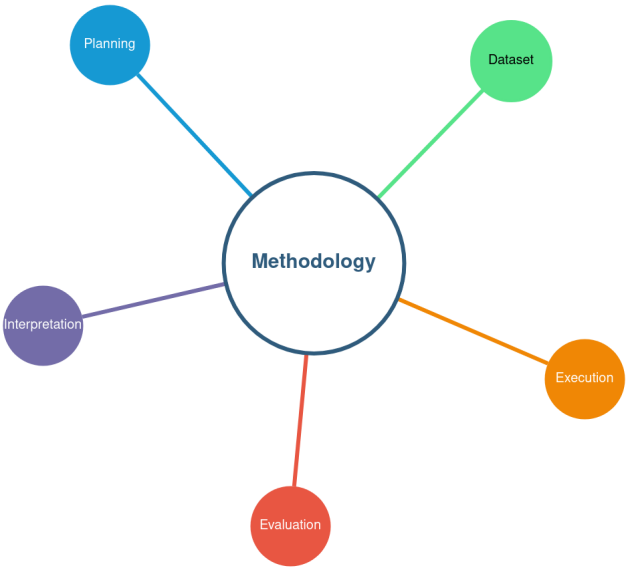


Figure 6.1: Methodology brief map

6.1 Planning phase

This section summarizes the main actions to take during planning the testing process. First part talks about identifying the system properties – logging the transactions, logging the results, system feedback, template updating and required input properties.

Second part talks about defining the system use-cases and the impact of system use-case on the testing process.

Third part is about the attacker model, the reason behind modeling and general practices. This section will tak mostly about the OWASP threat modeling [4] approach.

Fourth part discusses the goal of testing – why and when to set any goal and what impact does the goal have on further testing steps.

The last part deals with constructing a complete testing scenario using information from previous parts.

6.1.1 Identifying system properties

Upcoming part discusses important properties of the target of evaluation. Those properties come mainly from existing standards. At last there is a mention about the required system input as some biometric systems may require specific input properties that need to be addressed later (in the connection to dataset used).

Logging transactions

One of the standard-based parameters is logging the transactions. Every transaction made has to be logged to be able to trace and reproduce every step of the test. There are essentially two ways of logging the transactions:

- Automatic – the system itself keeps information about processed transactions either in database or in temporary cache. In case of caching the transactions, the tester needs to save them before the caches are cleared.
- Manual – the transactions are logged manually by the tester. In this case the tester must note all the relevant information (test subject, input data, ...).

Logging the results

Very similar to logging the transactions is logging the actual results of the tests. Every result must be logged either by system or by tester himself. Naturaly, every result must be logged to achieve realistic metrics. Again, there are two ways of looking at the logging of results:

- Automatic – the system itself keeps results either in database or in temporary cache. In case of caching the results, the tester needs to save them before the caches are cleared.
- Manual – the results are returned by system and need to be logged manually by the tester.

System feedback

The next property to be aware of is system feedback and how does to system communicate. This is an important information for the metrics selection and customization. There are two major options of system feedback:

- Score – the target of evaluation returns a number representing the score of sample evaluation. The score is usually either on a scale of 0 to 1 or -∞ to ∞ . The score typically requires either evaluating the results in a form of ROC/DET curves or setting a threshold(s) (more about setting a provisional threshold in 6.1.4.
- Accept/Reject – the target of evaluation returns predefined values symbolising either accepting or rejecting the input. Some systems also support the third state – something between accept and reject for when the matching algorithm is not sure. As for evaluation, existing thresholds allows tester to focus on metrics requiring accept/reject output (FAR, FRR, FMR, FNMR, ...).

Template updating

Template updating is a technique of continuous adaptation of user templates stored in the database based on the accepted attempts. Template updating is used to keep user templates up to date in case of dynamic biometrics that tend to change in time (signature, voice, thermogram, ...).

Template updating can e either manual – the staff manually adds new user samples to the system let it recompute the saved template, or automatic – the template is periodically updated with sampled input data that were accepted by the system. The nature of this approach makes testing the system's performance much more complex task 6.2, since order of data could reflect into the results.

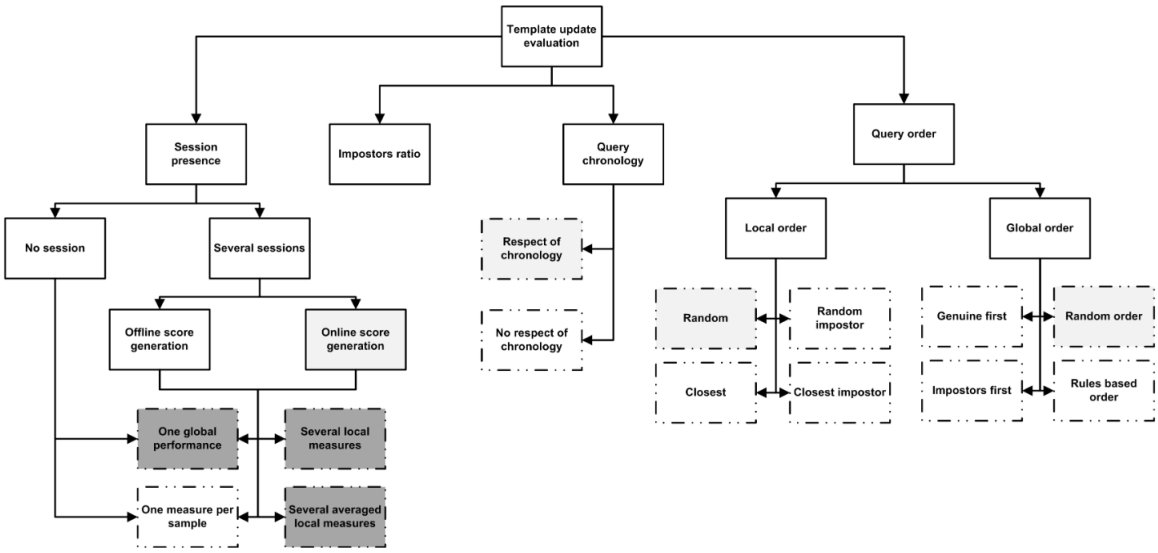


Figure 6.2: Template updating evaluation summary. Image retrieved from [15].

Required input properties

At last, the required input properties of a biometric system. This is a aspect that mostly affects the dataset selection phase. The required input data properties are dependent on the biometric the system is built to observe. Because this thesis is focused on deepfakes and deepfakes are mostly audio/video, I will only describe the potential properties of those two data formats.

- Audio/Video quality – audio/video quality could sometimes cause interesting results, when it comes to system robustness. Noisy audio recording or audio files with low sampling frequency could be rejected by the biometric system as too noisy to be evaluated. Same goes for low-quality/noisy video of videos with low resolution. For this reason, it is necessary to consider this factor when gathering the data or selecting the appropriate dataset.

- Audio/Speech/Video length – audio/speech length is a important aspect when it comes to performance. Voice biometric systems typically have a strict minimum of speech needed to enroll new user or to verify/identify a voice recording. Any shorter recording could be either rejected or evaluated as non-match. Again, same goes for video length.

- Speech/Video properties – other properties of the audio itself or the speaker. These are, for example, the language, sex of the speaker or age. Other features are highly dependent on the data acquisition phase, as existing datasets mostly do not list these features. Video properties to consider are the background of the subject or environment features.

6.1.2 Defining system use-cases

The next part discusses the system use-case selection. This part highly depends on the goal of testing. In case of technology testing, the use-case is irrelevant. But on the other hand, in case of scenario testing, the use-case of system could, and probably would, have impact on the data selection, test execution and even evaluation.

As mentioned in 4.1.3, for scenario testing the use-cases are divided into the groups according to the main function of a system (mode of comparison):

- verification – comparing the input samples and stored templates 1:1 – the user declares who he/she is and then presents the proof in the form of biometric

- identification – comparing the input samples and stored templates 1:M – the user does not declare who he/she is and presents the biometric, the system then tries to identify the user by matching the input sample to the existing templates in its database

The difference between them, in terms of dataset, is the data corpus layout. The idea of verification testing is to get verified as target user using the spoofed samples. But in case of identification, there could be two different goals – either to get identified as someone different or don't get identified at all. And this exact goal needs to affect the used dataset in a way, that the data spoof samples are in overlap with the registered users (always true for verification, as well as for attempting to be identified as target user) or the spoofs originates from unregistered users (avoiding identification).

A separate section belongs to the analysis systems. Even though this work didn't consider this general use-case, it is assumed to be similar to identification/verification testing.

The last mentioned case is the usage as integrated part of other products. This requires deep understanding of the outer system and all of it's parts that comes in touch with input biometric data. Any of these systems can preprocess the data, which needs to be remembered especially during the data acquisition.

6.1.3 Defining the attacker model

The attacker model is a way of better understanding the situation and existing threats to the target of evaluation. In general, there are multiple ways of evaluating the attacker/threat model. I this work, I decided to focus on two of these – OWASP Attacker model and attack rating (as proposed in [36]). First technique is structured into three sections:

- Motive – identifying the possibly motives of the attacker. What could possibly motivate the attacker to set the system as a target? What valuables does the system guard? What would be the benefits of breaking into the guarded product?
- Means – what would be the tools needed to attack the system. Is it hard to acquire the tools? How much they cost?
- Opportunity – what opportunities does the attacker have to attack the system. Is the system publicly available? Is any necessary data publicly available? Is there a supervision over the system sensors?

The other mentioned technique is rating the attack and evaluating the system's resistance. During the process, multiple factors are rated according to tables and then summarized into the overall system score. The detailed procedure is described in 4.1.1.

6.1.4 Determining the goal

As for setting the goal of testing, there are no specifics to be said. Setting of the main goal is primarily done for a simple introduction to the issue, setting a definitive objective and a strategy to achieve this.

Thresholds

In some cases, the goal of the test can require a very specific scenario of using the system. Especially in the event of testing a deployed product that is already integrated into a existing infrastructure. During such evaluation, concrete values of the system's thresholds need to be specified and later used for computing the metrics.

The companies could sometimes have a set of recommended threshold values, however, these often tend to be bound to a particular dataset. When testing the system using different data with predefined thresholds, the results could be misleading. In case of testing system with no specific thresholds set, I would recommend evaluating the system using just ROC/DET curves, as they capture the whole system's characteristic.

6.1.5 Defining the testing scenario

The last step of the planning phase is to summarize the previous information into a coherent test scenario. According to ISO/IEC 19795-2:2007 [17], a test scenario can be formulated

in two ways: technology testing or scenario testing. Each approach brings with it certain aspects of the test flow. These recommended features are briefly summarised in the table 6.1.

	Technology	Scenario
Target of evaluation	Biometric component (algorithm)	Biometric system
Goal	Algorithm performance evaluation	System performance evaluation (with simulated application)
Fundamental truth	Known association between source of data and samples	Known association between system decisions and sources of presented samples
Subject behaviour	Unusable	Directed
Real-time feedback	No	Yes
Repeatability	Yes	Partially (depends on data)
Environment oversight	Directed	Directed/Recorded
Interaction logging	Unusable	Recorded
Report with typical results	Relative robustness of components	Relative robustness of system
Typical metrics	Most error rates	Predicted end device throughput, FAR/FRR, FTA, FTE
Limitations	Suitable database	System deployment
Human subjects	Recorded	Live participation

Table 6.1: Technology vs Scenario summary

6.2 Dataset

The upcoming section discusses the most important part of the methodology for testing authentication biometric systems against deepfakes – the choice of datasets to use. The section covers both the use of existing datasets and the collection and use of custom datasets.

6.2.1 Using the existing datasets

When using publicly available online datasets, several considerations must be taken into account. The first, obvious, consideration is the composition of the dataset. It is not uncommon for the available datasets to consist of original (genuine) recordings, fakes created using deepfake technologies, as well as conventional methods. In these cases, it is necessary to have sufficient information about the dataset to distinguish between these parts, primarily the deepfake and conventional forgeries, especially when testing the system specifically against deepfake attacks. In the case of general system robustness testing, there is no need to record this distribution.

As with the distribution of genuine data and fakes, other aspects of the dataset must be considered according to the stated testing objective. In the case of monitoring other

characteristics, for example to test the robustness of different components or system functionalities, attention should be paid to the information provided about these characteristics. As outlined earlier in chapter 4.2, these characteristics are for example speech language, gender of subjects or age. The inclusion of this information is not common in public datasets, so it is necessary to verify this information before choosing a dataset.

Another very often mentioned data property is quality. More specifically, given the topic of this thesis, I am talking about audio quality. Audio quality is a very popular aspect of dataset creators. Commonly available datasets containing deepfake forgeries often include, in addition to clean audio samples, ones intentionally tainted either by artificial noise (added to existing samples from noise recording databases) or artificially caused, for example, by playing back samples using low-quality playback and recording equipment. Some biometric systems may be sensitive to the audio quality of the samples, so it is necessary to take this into account and either separate the samples during testing or note this fact during evaluation. However, sometimes this feature is desirable, for example for testing system filters or components that evaluate the quality of the input audio prior to the actual sample matching process. Again, it all depends on the stated testing objective.

The last thing that must be taken into account not only when selecting a dataset for testing, but also subsequently when interpreting the results is the target clientele of the tested system and its differences from the selected dataset. When using the available resources, it is typically not possible to select a perfect test suite in this regard. It will always differ from the target in some areas. Therefore, it is necessary to select, if possible, datasets that will not differ significantly in important, observed aspects and, on the contrary, unmeasured properties can be neglected. Nevertheless, it is important to be aware of this fact and to note it when interpreting the results.

Type	Name
Voice	ASVspoofFAD, , SV2TTS, WaveFake, SYNSPEECHDDB, FoR, FMFCC-A
Face	FakeAVCeleb, Celeb-DF, KoDF, DFDC, DeepFake MNIST+

Table 6.2: Existing datasets examples for voice and face deepfakes

6.2.2 Creating your own dataset

Collecting and creating custom datasets comes with many challenges. As this thesis focuses on the use of online available datasets and not on creating your own, the following section is just a brief overview of ideas and recommendations for data collection. The creation of custom datasets is addressed in other works.

Samples collection

The first step in collecting your own data is to determine the characteristics of your subject group. Given the available options, it is possible to customize the group of people according to the expected client base. As indicated earlier, in this regard we have the possibility to focus on important characteristics such as language, age, gender, as well as other aspects of the intended use of the system.

We can also include here the desired characteristics of the input data of the product under test – for example, the length of the recordings. At the same time, we have full control over the total length of the audio or the total amount of speech in each recording.

Another advantage is the precise control over data quality. We can include clearly defined scenarios of system operation in noisy environments and reflect this in the test dataset. We can also test system use-cases where users can also register from noisy environments – i.e., include a set of noisy genuine samples.

The quality of the audio is linked to the quality of the deepfake spoofs. As has been mentioned several times, the quality of deepfake fakes is a current subject of study and there is not yet a precise procedure to determine it. However, we can assume that it is possible to create deepfakes of different qualities that will affect the decision making of the system (as shown by the results of the testing of the delivered system in this work). Therefore, it is always advisable when creating or maintaining a dataset to seek out and study the latest techniques for creating deepfakes and, if possible, include them in the methods used to generate deepfakes.

Once the previously mentioned characteristics are decided, it's time for the actual data collection. There are few approaches to it. First thing to consider is whether to use real human subjects to collect desired dataset, or collect data samples on the public social sites, for example on YouTube[1] according to the terms of use.

This way, the approach to collection of data can be divided by method of creating the data:

- Real collection – fixed scenario of data collection using real human subjects.
- Using existing samples – variable (unsupervised) scenario of creating the samples.

Last thing to consider is the amount of data to collect. In general, the standards are very vague in terms of the amount of test samples in corpus – advice is to get as many as possible. Strictly speaking, the number of data samples does not matter in terms of conducting the test itself, but rather it affects the results relevance. For this purpose of statistical prove of relevance, there are two rules to follow when acquiring data – Rule of 3 and Rule of 30. Rule of 3 is about the smallest error rate while Rule of 30 is about the amount of data. Both of these rules are described in 6.5.

Forgeries synthesis

After collection of the genuine data corpus is done, it is time for synthesis of the forgeries part of arising dataset. There are numerous methods of creating a voice deepfake. As this work does not focus on creating the deepfake forgeries, but rather using them, I will not specify any of there. On this regard, there are other works that specializes in this field [14].

The important question is, how many of the deepfakes to synthesize? I have no definitive answer to this. A common ratio of bonafide samples and spoofs is around 1:10. But, in terms of separate sources of forgeries (a.k.a. the methods of generating deepfakes) I would suggest 1:1 bonafide to spoofs ratio.

6.3 Testing process

The next step after finishing up the planning of the test and selecting appropriate dataset is conducting the test. The process of conducting the test isn't very special and thus there are no steps or rules to follow. Depending on previous decisions and gathered information, however, there are some recommendations.

[1] https://www.youtube.com/

The testing process highly depends on the proposed testing scenario. But, both in case of testing the technology or testing the system, we need means of communication with the target of evaluation – a tool that ensures constant environment conditions and meets the requirements set in part about identifying system properties 6.1.1 of planning phase. Besides that, I would recommend implementing logging features to the testing tools regardless of the tested system properties.

6.4 Evaluation

After testing the system, we have records of individual transactions and system evaluation results. Now we just need to convert the measured results into metrics for evaluating the biometric system.

Again, the measurement metrics depend on the planned scenario. If only the technology is measured, i.e. the algorithm, standard FMR/FNMR metrics can be used for the case where we have clearly defined thresholds. If the thresholds are not known or clearly established, the results can be plotted in ROC/DET curves and then compared using AUC. All these metrics are described in 3.3. Other commonly used accuracy[11] and precision[32] metrics can be used to test the detection methods of the system.

In the case of scenario testing, other metrics introduced in the standard can be used depending on the situation. If sensors are used during system testing, FTA and FTE metrics can be also included. By using these metrics we also get the more commonly presented FAR/FRR metrics. Again, the same as for technology testing – if the exact threshold is not known, these metrics are plotted in ROC/DET curves and then compared using AUC.

As shown in this work, it is possible to introduce custom enhanced metrics to observe other properties related to the system's discriminative power. These observed properties can typically be tied to specific measured values, i.e., FAR/FRR or AUC from their representation by an ROC curve. The demonstration metric presented in this work was AUC vs. deepfake type.

6.5 Results interpretation

This brings us to the last section of this chapter – the relevance of the measured results. As indicated earlier, the significance of the results depends on the amount of data used. There are no exact numbers for determining the size of the test corpus. However, the standard does establish two rules on this topic:

- Rule 3 – sets statistically smallest error rate that can be set based on N independent comparisons – error rate p, when the probability of no error in N comparisons is 5%. This leads to $p \approx 3/N$ with the confidence of 95% [16].
- Rule 30 – tells us whether we used enough data. To be 90% sure that the true error rate lies within ±30% of the observed error rate, at least 30 errors must occur. [16]

Chapter 7

Conclusion

We live in the information era – era controlled by media and digital content. Social sites are an inseparable part of our everyday life. They contain news, videos, text posts, images or voice recordings. Social sites play not only the role of the communication medium, but also a role of a journalist platform with extremely wide reach in the society. Well targeted falsified messages have a potential to cause immeasurable damages.

Deepfakes, a. k. a. fake media generated using deep neural networks dominate the social networks today, which, among other things, often fulfills the function of news media. People are fascinated by the endless possibilities of the newly arrived technology. They entertain themselves by creating fun content, videos or believable parodies. Other than that the movie creators dream of the possibilities of using the deepfakes, especially in terms of filming a movie with actors, who are not able to perform anymore due various reasons. However on the other side stand the people, who have different, malicious plans with such powerful tools. Whether it is mass manipulation, blackmailing with highly targeted content, forgery of evidence or impersonating another person.

In order to protect people spending their time in the online space, or just casual users of the online services, new methods of deepfake detection are being developed. However, with the problem of detecting the deepfakes deals not only the common media, but so do the developers of biometric security systems.

Authentic biometrics of individuals can easily be generated the same way as the artificial fun videos and pictures but for the sole purpose of deceiving a access control system based specifically on those biometric features. Whether it is a picture of a face, a face recording or voice recording, which are very common and publicly accessible parts of a social network profile.

The biometric systems developers need to react quickly to the arrival of such powerful and accessible tools. Implementation of existing, proposed detection methods is not the only thing that needs to be focused on. Very important part of the development of biometric systems is testing.

Given the problem of falsifying biometric data, which has been a critical subject since the first biometric systems, the standards for unified testing have been established. They contain the suggested methodologies and recommendations to stick to when testing the developed system. These standards are sadly usually developed for many years, thus not considering the newest technologies, such as deepfakes, and their great influence on the detection systems.

The subject of this thesis is the problem of testing the robustness of a system against deepfakes. The main goal is a proposal of a general methodology, based on the current, well

known procedures, which focuses on the missing parts of testing the presentation attacks using forgeries generated by the modern deepfake technology. The methodology represents not only the general procedure, based as closely as possible on the current standards ISO/IEC, for testing the robustness of biometric systems against the spoofing attacks using deepfakes, but rather also advice and recommendations on which aspects to focus on and which not to neglect.

For the purpose of the methodology formulation and demonstrating the procedure I have proposed my own testing method of a supplied, commercially used voice biometric system Phonexia. The method is based on the standard-recommended practices with the addition of the focus on using publicly available deepfake data sets and proposing non-standard metrics as an example of possible monitoring of the various types of deepfakes according to their creation methods.

According to the proposed method, the testing of the system, as a tool used for remote voice-as-a-password verification, has been conducted. The detailed procedure and results are listed in the experiments chapter.

Main contribution of the thesis can therefore be summarized as a study and extension of current standard practices of biometric system testing by the area of testing the presentation attack using the modern forgeries generated using the deepfake technology. Demonstration and documentation of the method proposal for such testing step by step using the online, publicly available datasets. The most valuable part is, however, the methodology, as a generic repeatable way of testing the biometric systems that focuses on today's problems of deepfakes, based on the standard, proofed and well-known practices.

This work opens up many other directions for research in this area. One of them, already investigated and also mentioned in the thesis, is the quality of deepfakes. Being able to compare the quality of individual deepfakes would open up possibilities for more efficient development and testing of new defenses.

Another topic related to this thesis is the investigation of properties that affect the creation of deepfakes or the recognition of deepfakes by biometric systems. These are both content properties, hence for example speaker or speech properties, or properties of the recording itself.

Last but not least, there is the issue of the ever-increasing methods of generating deepfakes, which people find less and less recognizable. Research in this direction could show whether the same is true for biometric systems and thus point in the direction of evaluating generation methods associated with the quality of deepfakes.

A final topic is the real-time generation of deepfakes. As it is difficult, if not impossible, to implement an automatic liveness detection system in voice biometric systems, as mentioned in this work, it is often replaced by a dialogue during which the user is asked questions to which he/she must respond. Thus, it is assumed that the attacker must have pre-prepared samples to send to the system and that he will not be able to generate new samples during the conversation. If real-time deepfake generation systems emerge, this approach will also have to be abandoned and new protection options will have to be explored.

Bibliography

[1] *International Face Performance Conference* [online]. 2022 [cit. 2023-01-06]. Available at: https://www.nist.gov/news-events/events/2022/11/international-face-performance-conference-ifpc-2022.

[2] *Voice Verification & Speech Recognition Software* [online]. 2022 [cit. 2022-12-31]. Available at: https://www.phonexia.com/.

[3] *Classification: ROC Curve and AUC* [online]. 2023 [cit. 2023-05-07]. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[4] *OWASP DevSecOps Guideline* [online]. 2023 [cit. 2023-04-22]. Available at: https://owasp.org/www-project-devsecops-guideline/latest/00b-Threat-modeling.

[5] *Phonexia Speech Engine API Documentation* [online]. 2023 [cit. 2023-05-05]. Available at: https://download.phonexia.com/docs/spe/.

[6] Allyn, B. *Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn* [online]. 2022 [cit. 2022-12-22]. Available at: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[7] Bateman, J. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Carnegie Endowment for International Peace, 2020. Available at: http://www.jstor.org/stable/resrep25783.

[8] Bregler, C., Covell, M. and Slaney, M. Video Rewrite: Driving Visual Speech with Audio. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. USA: ACM Press/Addison-Wesley Publishing Co., 1997, p. 353–360. SIGGRAPH '97. DOI: 10.1145/258734.258880. ISBN 0897918967. Available at: https://doi.org/10.1145/258734.258880.

[9] Breland, A. *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink* [online]. 2019 [cit. 2022-12-22]. Available at: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

[10] Cabana, A., Charrier, C. and Louis, A. Mono and multi-modal biometric systems assessment by a common black box testing framework. *Future Generation Computer Systems*. 2019, vol. 101, p. 293–303. DOI: https://doi.org/10.1016/j.future.2019.04.053. ISSN 0167-739X. Available at: https://www.sciencedirect.com/science/article/pii/S0167739X1833111X.

[11] Chai, L., Bau, D., Lim, S.-N. and Isola, P. *What makes fake images detectable? Understanding properties that generalize*. 2020.

[12] Dack, S. *Deep Fakes, Fake News, and What Comes Next* [online]. 2019 [cit. 2022-12-22]. Available at: https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/.

[13] FBI. *Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions* [online]. 2022 [cit. 2022-12-22]. Available at: https://www.ic3.gov/Media/Y2022/PSA220628.

[14] Firc, A. *Applicability of Deepfakes in the Field of Cyber Security*. Brno, CZ, 2021. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Available at: https://www.fit.vut.cz/study/thesis/23761/.

[15] Giot, R., Rosenberger, C. and Dorizzi, B. *Performance Evaluation of Biometric Template Update*. 2012.

[16] *ISO/IEC 19795-1:2006 – Biometric performance testing and reporting — Part 1: Principles and framework*. Standard. International Organization for Standardization, april 2006.

[17] *ISO/IEC 19795-2:2007 – Biometric performance testing and reporting – Part 2: Testing methodologies for technology and scenario evaluation*. Standard. International Organization for Standardization, february 2007.

[18] *ISO/IEC 30107-3:2017 – Biometric presentation attack detection — Part 3: Testing and reporting*. Standard. International Organization for Standardization, september 2017.

[19] Jain, R. and Kant, C. Attacks on Biometric Systems: An Overview. *International Journal of Advances in Scientific Research*. september 2015, vol. 1, p. 283. DOI: 10.7439/ijasr.v1i7.1975.

[20] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J. et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *CoRR*. 2018, abs/1806.04558. Available at: http://arxiv.org/abs/1806.04558.

[21] Li, Y. and Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. *CoRR*. 2018, abs/1811.00656. Available at: http://arxiv.org/abs/1811.00656.

[22] Lin, C., Deng, J., Hu, P., Shen, C., Wang, Q. et al. *Towards Benchmarking and Evaluating Deepfake Detection*. arXiv, 2022. DOI: 10.48550/ARXIV.2203.02115. Available at: https://arxiv.org/abs/2203.02115.

[23] Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H. et al. *ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild*. arXiv, 2022. DOI: 10.48550/ARXIV.2210.02437. Available at: https://arxiv.org/abs/2210.02437.

[24] Ma, H. and Yi, J. *FAD: A Chinese Dataset for Fake Audio Detection*. Zenodo, june 2022. DOI: 10.5281/zenodo.6641573. Available at: https://doi.org/10.5281/zenodo.6641573.

[25] Mak, T. and Temple Raston, D. *Where Are The Deepfakes In This Presidential Election?* [online]. 2020 [cit. 2022-12-22]. Available at: https://www.npr.org/2020/10/01/918223033/where-are-the-deepfakes-in-this-presidential-election?t=1609838586916.

[26] Maras, M.-H. and Alexandrou, A. *Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos.* The International Journal of Evidence & Proof, 2018 [cit. 2022-12-22]. Available at: https://doi.org/10.1177/1365712718807226.

[27] Mirsky, Y. and Lee, W. The Creation and Detection of Deepfakes: A Survey. *CoRR.* 2020, abs/2004.11138. Available at: https://arxiv.org/abs/2004.11138.

[28] Navarro, F. *Deepfake porn videos are now being used to publicly harass ordinary people* [online]. 2019 [cit. 2022-12-22]. Available at: https://www.komando.com/security-privacy/deepfake-porn-videos-are-now-being-used-to-publicly-harass-ordinary-people/526877/.

[29] Qian, K., Zhang, Y., Chang, S., Yang, X. and Hasegawa Johnson, M. *AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss.* arXiv, 2019. DOI: 10.48550/ARXIV.1905.05879. Available at: https://arxiv.org/abs/1905.05879.

[30] Qian, K., Zhang, Y., Chang, S., Yang, X. and Hasegawa Johnson, M. *AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss.* 2019.

[31] Rathgeb, C., Tolosana, R., Vera Rodriguez, R. and Busch, C., ed. *Handbook of Digital Face Manipulation and Detection.* 1st ed. Springer Cham, 2022. 463-481 p. ISBN 978-3-030-87664-7.

[32] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. et al. *FaceForensics++: Learning to Detect Manipulated Facial Images.* 2019.

[33] Stewart, J. Q. An Electrical Analogue of the Vocal Organs. *Nature.* 09/1922 1922, p. 311–312. DOI: https://doi.org/10.1038/110311a0. Available at: https://www.nature.com/articles/110311a0#citeas.

[34] Takahashi, D. *McAfee shows how deepfakes can circumvent cybersecurity* [online]. 2019 [cit. 2022-12-22]. Available at: https://venturebeat.com/ai/mcafee-shows-how-deep-fakes-can-circumvent-cybersecurity/.

[35] Taylor, P. The text-to-speech problem. In: *Text-to-Speech Synthesis.* Cambridge University Press, 2009, p. 26–51. DOI: 10.1017/CBO9780511816338.005.

[36] Tekampe, N., Merle, A., Bringer, J., Gomez Barrero, M., Fierrez, J. et al. D6.5: Towards the Common Criteria evaluations of biometric systems. In: *BEAT: Biometrics Evaluation and Testing.* 2016.

[37] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C. and Niessner, M. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016, p. 2387–2395. DOI: 10.1109/CVPR.2016.262.

[38] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H. et al. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. arXiv. 2019. DOI: 10.48550/ARXIV.1904.05441. Available at: https://arxiv.org/abs/1904.05441.

[39] Unar, J., Seng, W. C. and Abbasi, A. A review of biometric technology along with trends and prospects. *Pattern Recognition*. 2014, vol. 47, no. 8, p. 2673–2688. DOI: https://doi.org/10.1016/j.patcog.2014.01.016. ISSN 0031-3203. Available at: https://www.sciencedirect.com/science/article/pii/S003132031400034X.

[40] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A. et al. *ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech*. 2020.

[41] Westerlund, M. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*. Ottawa: Talent First Network. 11/2019 2019, vol. 9, p. 40–53. DOI: http://doi.org/10.22215/timreview/1282. ISSN 1927-0321. Available at: timreview.ca/article/1282.

[42] Yang, X., Li, Y. and Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. *CoRR*. 2018, abs/1811.00661. Available at: http://arxiv.org/abs/1811.00661.

Appendix A

Media contents

```
+- paper    - directory with source code for this work
|
+- results  - directory with results file
|
+- tool     - directory containing the scripts used for testing and
    .         parsing the results
    .
    +- README.md    - contains the tool description and usage
```

Appendix B

Attack rating tables

Factor	Identification	Exploitation
Elapsed time		
<= one day	0	0
<= one week	1	2
<= two weeks	2	4
<= one month	4	8
>= one month	8	16
Expertise		
Layman	0	0
Proficient	2	4
Expert	4	8
Multiple Experts	8	0 (Not applicable)
Knowledge of system		
Public	0	0 (Not applicable)
Restricted	2	0 (Not applicable)
Sensitive	4	0 (Not applicable)
Critical	8	0 (Not applicable)
Access to the system / Window of opportunity		
Easy	0	0
Moderate	2	4
Difficult	4	8
Equipment		
Standard	0	0
Specialized	2	4
Bespoke	4	8
Access to biometric characteristics		
Immediate	0 (Not applicable)	0
Easy	0 (Not applicable)	2
Moderate	0 (Not applicable)	4
Difficult	0 (Not applicable)	8

Table B.1: Rating table. Retrieved from [36].

Value	Biometrics modality
Immediate	2D Face, Signature Image, Speech
Easy	Fingerprint
Moderate	Iris, 3D Face, Dynamic Signature, 3D Fingerprint
Difficult	Veins

Table B.2: Rating for biometric modalities. Obtained from [36]

Values	Attack potential for the whole attack	System resistant to attackers with attack potential of
<10	Basic	No rating
10-19	Enhanced-Basic	Basic
20-29	Moderate	Enhanced-Basic
30-39	High	Moderate
>=40	Beyond high	High

Table B.3: Resistance levels table. Table retrieved from [36]